**PROJECT-2**

**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.

**2. Предварительный анализ данных**

**Задание 2.1**

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

**Код запроса:**

SELECT

max(age) as max\_age -- применим агрегирующую функцию к столбцу возраст

FROM

hh.candidate

**Результат: 100**

**Вывод:** Можно было догадаться, что результат будет именно таким. Ведь мы исследуем тот же датасет что и в Project-1. Соискатель с таким возрастом – это аномальное значение, думаю трудно сохранять работоспособность в данном возрасте.

**Задание 2.2**

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

**Код запроса:**

SELECT

min(age) as min\_age -- применим агрегирующую функцию к столбцу возраст

FROM hh.candidate

**Результат: 14**

**Вывод:** Тут уже более реальная картина, так как по трудовому законодательству РФ устроиться на работу можно с 14 лет, с согласия родителей. Проанализировав датасет с сортировкой по возрасту, можем увидеть, что у данного кандидата занимаемая и желаемая должность одна и та же, что скорее всего опечатка. Кандидаты в возрасте от 15 до 18 так или иначе связанны с IT сферой, что вполне реально и очень востребовано в наше время.

**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.

**Код запроса:**

SELECT

age, -- столбец возраст

COUNT(\*) cnt -- счетчик для каждого возраста

FROM hh.candidate

GROUP BY 1 -- группировка по возрасту

ORDER BY 1 DESC -- сортировка по возрасту в порядке убывания

**Результат:**

**Изображение выглядит как стол

Автоматически созданное описание**

**Вывод:** В результате запроса мы видим, что кандидат с возрастом 100 лет встречается 1 раз, что вполне логично, в дальнейшем мы не будет учитывать его в наших исследованиях. По остальным кандидатам количество вполне реалистичное. Если же посмотреть на сортировку в порядке возрастания, то хорошо видно, что больше всего соискателей в возрастной группе от 22 до 39 лет, дальше количество заметно снижается.

**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

COUNT(id) -- счетчик кандидатов

FROM hh.candidate

WHERE age > 40 AND age != 100 -- условие, возраст больше 40 лет и не равен 100

**Результат: 6,263**

**Вывод:** В результате запроса получаем 6263 соискателя, которые старше 40 лет. Что примерно равно 14% от общего количества соискателей. Как правило данная категория соискателей уже имеет хорошую должность и опыт работы.

**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.

**Код запроса:**

SELECT

c.title as city, -- столбец с названием городов, алиас city

COUNT(\*) as cnt -- счетчик количества кандидатов в каждом городе

FROM

hh.candidate as can

JOIN hh.city as c ON c.id=can.city\_id -- соединение таблицы кандидатов с таблицей городов

GROUP BY city -- группировка по столбцу city

ORDER BY 2 DESC -- группировка по столбцу количества в порядке убывания

**Результат:**

****

**Вывод:** Из таблицы видно, что почти половина соискателей ищет работу в городах-миллионниках. Это вполне нормальное явление, так как уровень жизни, доходы населения в этих городах выше.

**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

gender, -- столбец пол

age, -- столбец возраст

desirable\_occupation, -- столбец желаемая должность

c.title city, -- столбец город

ca.employment\_type -- столбец вид занятости

FROM

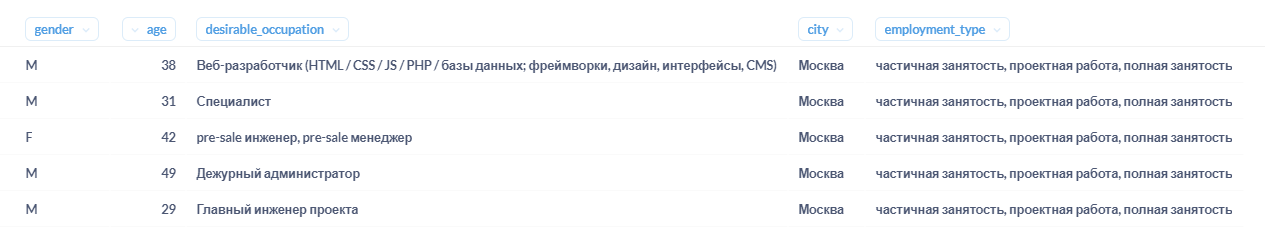
hh.candidate ca

JOIN hh.city c ON ca.city\_id = c.id -- соединение таблицы кандидатов с таблицей города

WHERE c.title LIKE 'Москва' and ca.employment\_type LIKE '%проектная работа%' -- условие, интересующий нас город – Москва и кандидат не исключает занятости «проектная работа»

ORDER BY ca.id -- сортировка по id кандидата

**Результат:**

****

**Вывод:** Если более детально изучить данные, то можно заметить, что таких соискателей 2950. Это примерно 18% от всех кандидатов, ищущих работу в Москве. Наиболее активная возрастная группа от 21 до 34 лет. Скорее всего это кандидаты, только закончившие колледж и институт, а также начинающие свою трудовую карьеру.

**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

ca.gender, -- столбец пол

ca.age, -- столбец возраст

ca.desirable\_occupation, -- столбец желаемая должность

c.title city, -- столбец город

ca.employment\_type -- столбец вид занятости

FROM

hh.candidate as ca

JOIN hh.city as c ON ca.city\_id = c.id -- соединение таблицы кандидатов с таблицей города

WHERE c.title = 'Москва'

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

AND (lower(ca.desirable\_occupation) LIKE '%разработчик%'

OR lower(ca.desirable\_occupation) LIKE '%аналитик%'

OR lower(ca.desirable\_occupation) LIKE '%программист%') -- условие, интересующий нас город – Москва и кандидат не исключает занятости «проектная работа», а также фильтр по профессиям разработчик, аналитик, программист

ORDER BY ca.id -- сортировка по id кандидата

**Результат:**

Изображение выглядит как стол

Автоматически созданное описание

**Вывод:** Посчитав количество кандидатов, удовлетворяющих условию из задания, получаем 778 человек, что составляет 26,5% от готовых к проектной работе. Самая распространенная профессия – программист, затем разработчик и замыкает тройку аналитик. Наиболее активная возрастная группа от 21 до 34 лет. Мне кажется, данные профессии склонны к исследовательским работам.

**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

ca.id, -- столбец id кандидата

c.title as city -- столбец город

FROM hh.candidate ca

JOIN hh.city c ON ca.city\_id = c.id -- соединение таблицы кандидатов с таблицей города

WHERE ca.desirable\_occupation = ca.current\_occupation -- условие, желаемая должность совпадает с занимаемой

ORDER BY city, ca.id – сортировка по городу, а затем по id кандидата

**Результат:**



**Вывод:** Посчитав данную группу людей получаем 5103 человека, что примерно 11% от общего числа кандидатов. Мое мнение, что кандидатам просто хочется сменить место работы, не меняя должность, возможно с повышением заработной платы. И снова, активная возрастная группа от 21 до 34 лет.

**Задание 3.5**

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

**Код запроса:**

SELECT

COUNT(id) -- счетчик

FROM hh.candidate

WHERE age != 100

AND (gender = 'M' AND age > 64) OR (gender = 'F' AND age > 59) – условие пенсионного возраста, также исключаем из запроса кандидата с возрастом 100 лет

**Результат: 75**

**Вывод:** Стоило ожидать такой результат, это примерно 0,17% от общего количества кандидатов, из них 57 мужчин и 18 женщин. Как говорил уже ранее, трудно сохранять работоспособность в пенсионном возрасте.

**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

ca.gender, -- столбец пол

ca.age, -- столбец возраст

ca.desirable\_occupation, -- столбец желаемая должность

c.title city, -- столбец город

ca.employment\_type, -- столбец вид занятости

tt.title timetable\_type -- столбец рабочий график

FROM hh.candidate ca

JOIN hh.city c ON c.id = ca.city\_id -- соединение таблицы кандидатов с таблицей города

JOIN hh.candidate\_timetable\_type ct ON candidate\_id = ca.id -- соединение дополнительной таблицы с таблицей кандидатов

JOIN hh.timetable\_type tt ON tt.id = ct.timetable\_id -- соединение таблицы-справочника с дополнительной таблицей

WHERE tt.title = 'вахтовый метод' AND c.title in ('Новосибирск', 'Омск', 'Томск', 'Тюмень') – условие, работа вахтовым методом и из данных городов

ORDER BY c.title, ca.id – сортировка по городу и по id кандидата

**Результат:**

Изображение выглядит как текст

Автоматически созданное описание

**Вывод:** Таких кандидатов всего 11 человек. Возможно, для данной компании стоит подыскать сотрудника на удаленную работу.

**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.

**Код запроса:**

(SELECT

ca.desirable\_occupation, -- столбец желаемая должность

ca.age -- столбец возраст

FROM hh.candidate ca

JOIN hh.city c ON ca.city\_id = c.id -- соединение таблицы кандидатов с таблицей города

WHERE c.title = 'Санкт-Петербург' AND ca.age BETWEEN 16 AND 21 – условие по городу и возрасту

ORDER BY ca.age -- сортировка по возрасту

LIMIT 10) -- лимит 10 кандидатов

UNION ALL -- объединение

SELECT

'Total', -- название итоговой строки

COUNT(ca.id) – общее количество кандидатов

FROM hh.candidate ca

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

WHERE c.title = 'Санкт-Петербург' AND ca.age BETWEEN 16 AND 21

**Результат:**

****

**Вывод:** Из запроса можно заметить, что ТОП-10 профессий связаны с IT сферой. Можно было предположить, так как IT сфера стремительно развивается.

**Общий вывод:** По результатам исследования датасета можно выделить несколько пунктов:

* 1. Тройка самых востребованных городов – Москва, Санкт-Петербург, Краснодар. Вакансий в данных городах больше, чем в остальных.
  2. Самая активная возрастная категория от 21 года до 34 лет, этого стоило ожидать, молодые специалисты всегда в спросе.
  3. Соискателей пенсионного возраста немного, но все же они есть.
  4. 43,284 готовы к полной занятости.