Проект 2.

Подгрузка новых данных. Уточнение анализа.

BY STANISLAV OLIINYK

*Описание проекта*

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. Работа с системами управления базами данных, возможность вывода информации и поиск статистических показателей являются одними с основополагающих навыков профессии Data Science.

*Какой кейс решаем?*

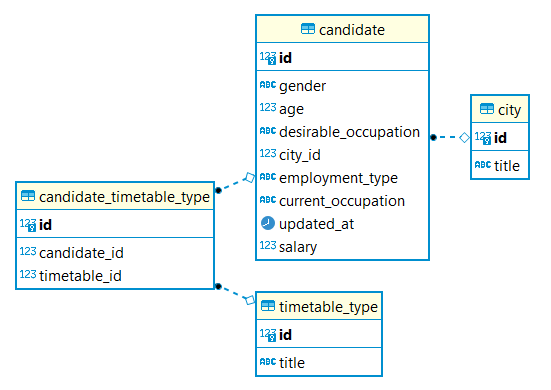
Необходимо познакомиться с данными, понять, с какими резюме вы имеете дело, а дальше собрать статистику для различных кадровых агентств и статистических центров.

*Что практикуем?*

Учимся работать с базами данных с помощью PostgreSQL.

*Дисклеймер*

Результаты запросов представлены в виде встроенных объектов OLE (англ. Object Linking and Embedding). С помощью двойного нажатия доступен просмотр полной таблицы.

1. Введение. Знакомство с датасетом

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 - 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.

TIMETABLE\_TYPE - это таблица-справочник вариантов рабочего графика, подходящего кандидату.

2. Предварительный анализ данных

Задание 2.1

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

SELECT /\*выбор\*/

MAX(age) max\_age /\*максимальное значение возраста с алиас max\_age\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

Результат: 100

Выводы:

Максимальный возраст кандидата 100 лет, что является выбросом данного показателя и может указывать на неверность указной даты рождения.

Задание 2.2

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

SELECT /\*выбор\*/

MIN(age) min\_age /\*минимальное значение возраста с алиас min\_age\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

Результат: 14

Выводы:

Минимальный возраст кандидата 14 лет, что обусловлено кандидатами с неполным средним образованием.

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.

SELECT /\*выбор\*/

age, /\*возраст\*/

COUNT(\*) cnt /\*подсчёт всех строк с алиас cnt\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

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

ORDER BY 1 DESC /\*сортировка по первому столбцу в порядке убывания\*/

Результат:



Выводы:

Большинство кандидатов с порогом от 15 находятся в возрасте от 18 до 65 лет, что соответствует трудоспособному возрасту. Мода соискателей приходиться на 30 лет в количестве 2834 заявок. Основная масса заявок (более 1000 на каждый возраст) находиться в интервале между 21 и 40 лет.

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.

SELECT /\*выбор\*/

COUNT(\*) /\*подсчёт всех строк\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

WHERE age BETWEEN 41 AND 99 /\*при условии, что возраст лежит в промежутке между 41 и 99\*/

Результат: 6263

Выводы:

Хотя средний возраст занятых в России составляет около 40 лет, количество кандидатов старше данного возраста составляет 6263, что 14% от всего количества. Это указывает, что тенденция поиска работы чаще присутствует у молодых людей. А люди после 40 продолжают карьеру на прежнем рабочем месте.

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. Отсортируйте результат по количеству в обратном порядке.

SELECT /\*выбор\*/

ci.title city, /\*столбец title таблицы ci с алиас city\*/

COUNT(\*) cnt /\*подсчёт всех строк с алиас cnt\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

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

ORDER BY 2 DESC /\*сортировка по второмустолбцу в порядке убывания\*/

Результат:



Выводы:

В выборке кандидатов в зависимости от города, как можно предвидеть, лидером по количеству заявок является город Москва. Столица значительно имеет отличимое значения открытых вакансий, а также и численность кандидатов по поиску работы. Второе место занимает Санкт-Петербург, как второй город по населению и значимости в России.

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.

SELECT /\*выбор\*/

ca.gender, /\*пол из таблицы ca\*/

ca.age, /\*возраст из таблицы ca\*/

ca.desirable\_occupation, /\*желаемая должность из таблицы ca\*/

ci.title city, /\*столбец title таблицы ci с алиас city\*/

ca.employment\_type /\*тип занятости из таблицы ca\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

WHERE ci.title = 'Москва' /\*при условии, что имя города столбец title из таблицы ci равен 'Москва'\*/

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

/\*и условии, что столбец типа занятости employment\_type из таблицы ca содержит «проектная работа»\*/

ORDER BY ca.id /\*сортировка по столбцу id из таблицы ca\*/

Результат:



Выводы:

В Москве присутствует 2851 кандидат готовый работать по такому типу трудоустройства, как проектная работа. Данный тип занятости в основном соответствует IT-профессиям и инженерному направлению.

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.

SELECT /\*выбор\*/

ca.gender, /\*пол из таблицы ca\*/

ca.age, /\*возраст из таблицы ca\*/

ca.desirable\_occupation, /\*желаемая должность из таблицы ca\*/

ci.title city, /\*столбец title таблицы ci с алиас city\*/

ca.employment\_type /\*тип занятости из таблицы ca\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

WHERE ci.title = 'Москва' /\*при условии, что имя города столбец title из таблицы ci равен 'Москва'\*/

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

/\*и условии, что столбец типа занятости employment\_type из таблицы ca содержит «проектная работа»\*/

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

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

OR LOWER(ca.desirable\_occupation) LIKE '%программист%')

/\*и условии, что столбец желаемой должности в нижнем регистре desirable\_occupation из таблицы ca содержит «разработчик» , «аналитик» или «программист»\*/

ORDER BY ca.id /\*сортировка по столбцу id из таблицы ca\*/

Результат:



Выводы:

779 кандидатов из Москвы готовы к проектной работе и являются обладателями самых популярных IT-профессий.

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.

SELECT /\*выбор\*/

ca.id id, /\*идентификатор из таблицы ca с алиас id\*/

ci.title city /\*столбец title таблицы ci с алиас city\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

WHERE ca.desirable\_occupation = ca.current\_occupation

/\*при условии, что желаемая должность desirable\_occupation равна занимаемой urrent\_occupation из таблицы ca\*/

ORDER BY 2, 1 /\*сортировка по второму и первому столбцам\*/

Результат:



Выводы:

Взглянув на данную таблицу результатов можно увидеть, что в Москве и Санкт-Петербурге больше всего кандидатов с совпадение желаемой и занимаемой должности. Это связано с количеством заявок. При более глубоком анализе можно сделать утверждение, что совпадение не зависит от города и равно 11%±1%.

Задание 3.5

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

SELECT /\*выбор\*/

COUNT(\*) /\*подсчёт всех строк\*/

FROM hh.candidate /\*из таблицы hh.candidate\*/

WHERE (gender = 'M' AND age BETWEEN 65 AND 99)

OR (gender = 'F' AND age BETWEEN 60 AND 99)

/\*при условии, что для мужского пола возраст лежит в промежутке между 65 и 99 лет;

или для женского - в промежутке между 60 и 99 лет \*/

Результат: 75

Выводы:

Количество кандидатов пенсионного является ничтожно малым для данной выборки в размере меньше 0,2% от суммарного количества записей.

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.

SELECT /\*выбор\*/

ca.gender, /\*пол из таблицы ca\*/

ca.age, /\*возраст из таблицы ca\*/

ca.desirable\_occupation, /\*желаемая должность из таблицы ca\*/

ci.title city, /\*столбец title таблицы ci с алиас city\*/

ca.employment\_type /\*тип занятости из таблицы ca\*/

t.title timetable\_type /\*тип рабочего графика из таблицы t\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

LEFT JOIN hh.candidate\_timetable\_type ct ON ca.id = ct.candidate\_id

/\*оператор левого соединения таблиц;

таблица hh.candidate\_timetable\_type с алиасом ct;

условие, что id таблицы ca равен candidate\_id таблицы ct\*/

LEFT JOIN hh.timetable\_type t ON ct.timetable\_id = t.id

/\*оператор левого соединения таблиц;

таблица hh. timetable\_type с алиасом t;

условие, что timetable\_id таблицы ct равен id таблицы t\*/

WHERE ci.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень')

/\*при условии, что название города title из таблицы ci

находиться в значениях 'Новосибирск', 'Омск', 'Томск' или 'Тюмень'\*/

AND lower(t.title) = 'вахтовый метод'

/\*и условии, что столбец рабочего графика из таблицы t

в нижнем регистре title соответствует вахтовому методу\*/

ORDER BY city, ca.id /\*сортировка по столбцу города city и id из таблицы ca\*/

Результат:

Выводы:

Под критерии заказчика подходят 12 кандидатов мужского пола в возрасте от 23 до 42 лет. Желаемая должность кандидатов варьируется и может удовлетворить более точный запрос по найму.

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.

( /\*открытие скобок\*/

SELECT /\*выбор\*/

ca.desirable\_occupation, /\*желаемая должность из таблицы ca\*/

ca.age, /\*возраст из таблицы ca\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci; условие, что city\_id таблицы ca равен id таблицы ci\*/

WHERE

ci.title = 'Санкт-Петербург' /\*при условии, что имя города столбец title из таблицы ci

равен «Санкт-Петербург»\*/

AND ca.age BETWEEN 16 AND 21

/\*и условии, что возраст лежит в промежутке между 16 и 21\*/

ORDER BY age LIMIT 10 /\*сортировка по возрасту и ограничение десятью значениями\*/

) /\*закрытие скобок для применения сортировки и ограничения

только для первой части запроса\*/

UNION ALL /\*оператор присоединения\*/

SELECT /\*выбор\*/

'Total', /\*сами задаём объект «Total»\*/

COUNT(\*) /\*подсчёт всех строк\*/

FROM hh.candidate ca /\*из таблицы hh.candidate с алиас ca\*/

LEFT JOIN hh.city ci ON ca.city\_id = ci.id

/\*оператор левого соединения таблиц; таблица hh.city с алиасом ci;

условие, что city\_id таблицы ca равен id таблицы ci\*/

WHERE

ci.title = 'Санкт-Петербург' /\*при условии, что имя города столбец title из таблицы ci

равен «Санкт-Петербург»\*/

AND ca.age BETWEEN 16 AND 21

/\*и условии, что возраст лежит в промежутке между 16 и 21\*/

Результат:



Выводы:

В таблице мы видим, что большинство молодых людей в Санкт-Петербурге выбираю IT-индустрию для будущей работы. В базе обнаружился 161 кандидат от 16 до 21 года. Дети знакомятся с компьютером с ранних лет и он остается важным элементом их жизни, что и определяет интерес к информационным технологиям.

Общий вывод по результатам анализа рынка труда.

Рынок труда является соотношением вакантных рабочих мест и незанятой рабочей силы. Трудовые ресурсы несут важную информация для понимания возможностей развития тех или иных направлений и открытые данные могут помочь с планированием и прогнозирование роста любой индустрии.

Выполняя анализ рынка труда было определено, что основная масса соискателей соответствует работоспособному возрасту по законодательству России от 16 до 65 лет. Больше всего заявок кандидатов 30-го возраста и основное число резюме приходиться на промежуток с 21 до 40 лет, что обусловлено выпускниками вузов и неопределенностью карьеры у молодого населения до 35 лет.

Москва, как столица и Санкт-Петербург охватывают самый большой сегмент кандидатов по причине присутствия существенного прироста открытых вакансий и новых мест работы.

С каждым годом IT-профессий становятся основными направлениями современного рынка за счет высокой заработной платы и удаленных типов занятости главным образом у специалистов до 21 года.

В данном анализе был выполнен подбор кандидатов для заказчика по указанным показателям и предоставлены данные для дальнейшего использования их при подборе, как вариантов на желаемые должности.