Документация

Дипломная работа по теме:
“Анализ рынка вакансий по данным сайта «Хабр Карьера» (парсинг данных, поиск инсайтов и составление рекомендаций стейкхолдерам)”

Профессия «Аналитик данных», DA-70
Летуновский Михаил Валериевич

г. Москва, 2023

Введение

Цели проекта:

Исследовать рынок вакансий в текущем моменте времени. Для анализа берутся данные с сайта Хабр.Карьера - тематический ресурс, на котором размещаются вакансии, связанные с информационными технология, аналитикой и IT-сферой в целом.

Бизнес-Задачи:

1. Изучить и выявить наиболее востребованные профессия на рынке труда в настоящий момент.

2. Выявить компании, которые в настоящий момент наиболее заинтересованы в найме сотрудником

3. Произвести анализ уровня заработной платы на рынке труда, в зависимости от навыков и опыта кандидата.

4. Определить наиболее востребованные навыки и знания, которые в настоящий момент востребованы и пользуются большим спросом.

5. Изучить географию предложений, в каких городах требуются специалисты.

Блок 1. Описание исходного датасета и типов данных (7 столбцов)

Для исследования данных, был написан парсер на языке Python, при помощи библиотек BeautifulSoup и request. В результате работы парсера была получена следующая информация с сайта https://career.habr.com/ :

- Название вакансии
- Название компании
- Место локации/формат работы
- Ссылка на вакансию
- Дата публикации объявления
- Навыки
- Заработная плата

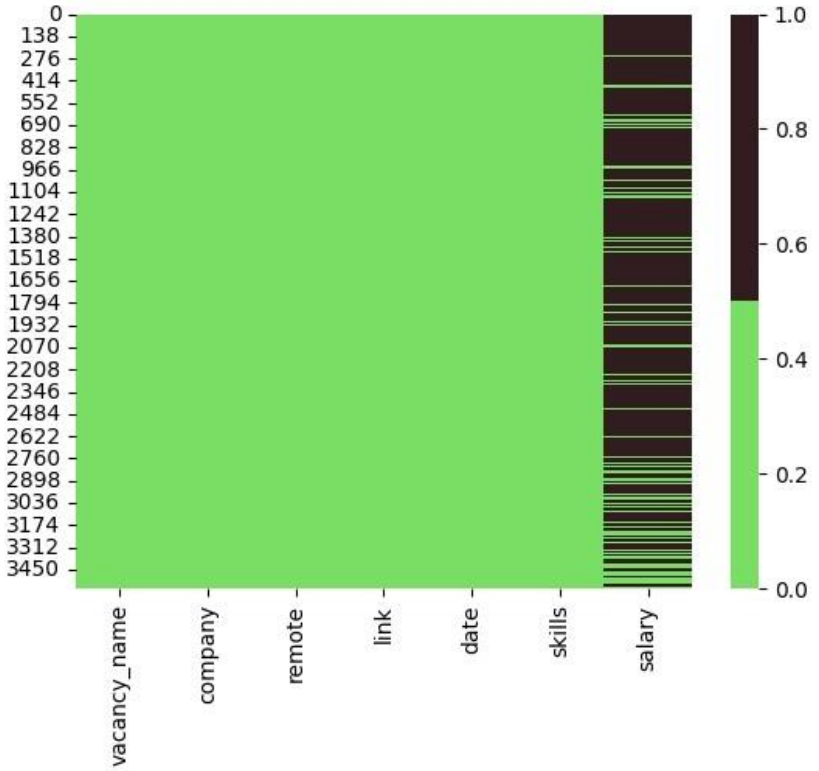
Далее при помощи библиотеки Pandas эти данные были преобразованы в табличную структуру.
В результате работы парсера, была сформирована таблица, состоящая из 3551 строки.

№	Имя Столбца	Описание	Тип данных
1	***vacancy_name***	Наименование вакансии	object
2	***company***	Название компании	object
3	***remote***	Город локации/тип занятости	object
4	***link***	Ссылка на вакансию	object
5	***date***	Дата публикации вакансии	object
6	***skills***	Необходимые навыки	object
7	***salary***	Уровень заработной платы	object

Блок 2. Подготовка и преобразование данных

2.1 Проверка на отсутствующие значения

В ходе исследования качества данных было проведено исследование на поиск пропущенных значений. В результате было установлено, что пропуске присутствуют только в одном столбце salary (уровень заработной платы). Уровень пропущенных значений составляет 82%.




```
vacancy_name - 0% пропущенных значений
company - 0% пропущенных значений
remote - 0% пропущенных значений
link - 0% пропущенных значений
date - 0% пропущенных значений
skills - 0% пропущенных значений
salary - 82% пропущенных значений
```

Вывод: в целом мы работаем с очень хорошим набором данных, отсутствующие значения есть только в одном столбце. Данная информация нам говорит о том, что работодатели не хотят сразу озвучивать предложения по зарплате, и скорее всего делают предложения соискателю исходя из опыта и навыков.

2.2 Преобразование данных

2.2.1 Преобразование данных в столбце remote

Как видно из набора значений - данные в каждой строке могут состоять из различных комбинаций названий городов или городов и типа занятости. Типы занятости представлены в трех возможных вариантах:

- Полный рабочий день
- Можно удаленно

- Неполный рабочий день

Поступим с данными следующим образом: для типов занятости создадим отдельный столбец kind_of_work и заполним его вышеуказанными тремя типами занятости, если в столбце remote указан только город, без указания типа занятости, то в этом случае в столбце kind_of_work появиться значения «Офис».

После этого из столбца remote удалим типы занятости, оставив только названия городов. Если город не будет указан, то отсутствующие значения заполним «Город не указан».

2.2.2 Преобразование данных в столбце date

Формат записи в столбце date - это «число месяц» и тип данных Object. Преобразуем его в удобный формат «число-месяц-год» при помощи библиотеки datetime.

2.2.3 Преобразование данных в столбце salary

В данном столбце 82% пропущенных значений и к тому же очень неудобная структура записи информации, а именно могут встречаться варианты " от сумма до сумма", "от сумма", "до сумма", "сумма". Зарплата может исчисляться в рублях, евро, долларах и тенге. Задача по очистке данных сводиться к следующим действиям:

1. Создаем два столбца "salary_from"(зарплата от) и "salary_to"(зарплата до).
2. Удаляем символы платежных валют и переводим значения в целочисленный формат.
3. Приводим значения в единую систему исчисления, а именно в рубли. Т.к. проект исследовательский, а текущий уровень инфляции очень большой, плюс большая волатильность на рынке валют, то курс доллара и евро будем считать 1 к 100 рублям, а курс тенге 1 к 4.58.

2.2.3 Итоговый датасет после преобразования данных

Выполнив преобразование данных, мы получили следующую таблицу:

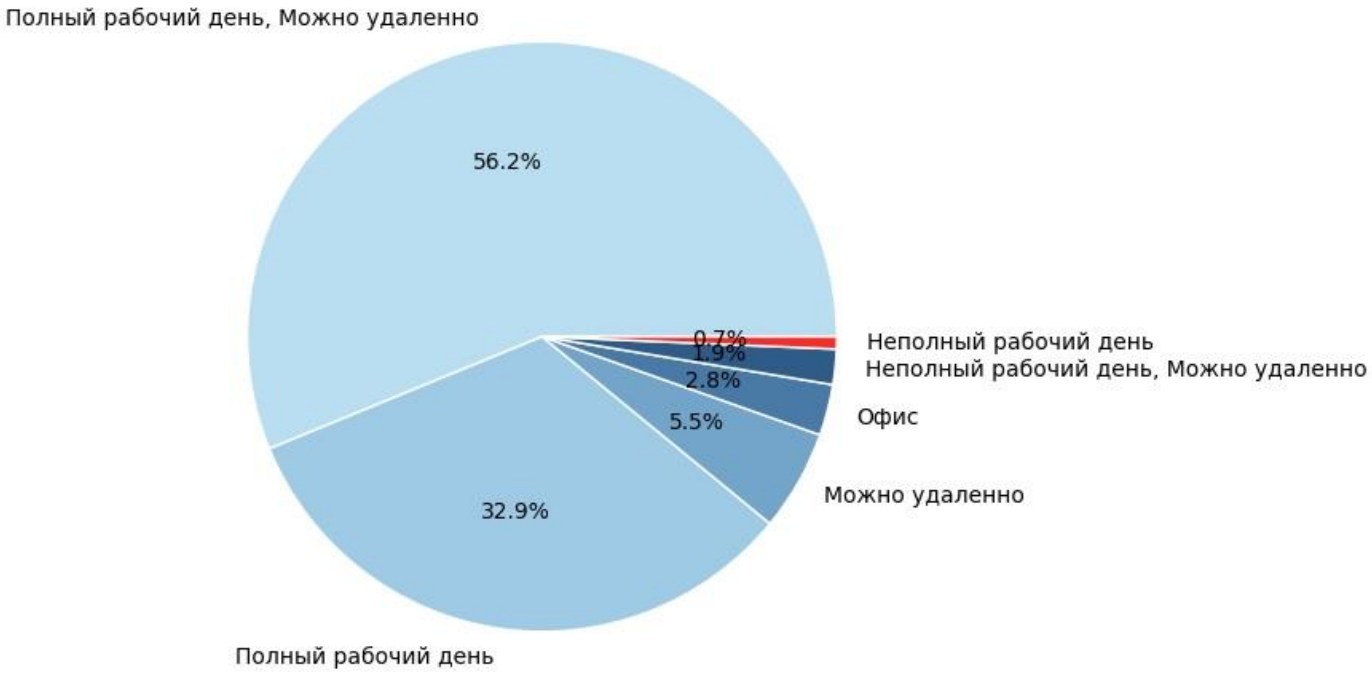
№	Новое имя Столбца	Преобразование данных	% NaN	Очистка данных
1	***vacancy_name***	object	0.00	Без изменений.
2	***company***	object	0.00	Без изменений.
3	***remote***	object	0.00	Удалены названия типов занятости, отсутствующие значение заменены на «Город не указан»
4	***link***	object	0.00	Без изменений.
5	***date***	datetime64[ns]	0.00	Дата преобразована в формат datetime «день-месяц-год»

6	***skills***	object	0.00	Без изменений.
7	***salary***	object	82	Без изменений. Пропуски сохранены
8	***kind_of_work***	object	0.00	В столбец добавлены типы занятости, пропущенные значения заменены на np.nan
9	***salary_from***	int64	85	В столбец добавлены значения «зарплата от», пропущенные значения заменены на np.nan
10	***salary_to***	int64	89	В столбец добавлены значения «зарплата до», пропущенные значения заменены на np.nan

Блок 3. Анализ данных для стейкхолдеров

Целью блока является систематизация и обобщение данных.

3.1 Анализ типов занятости

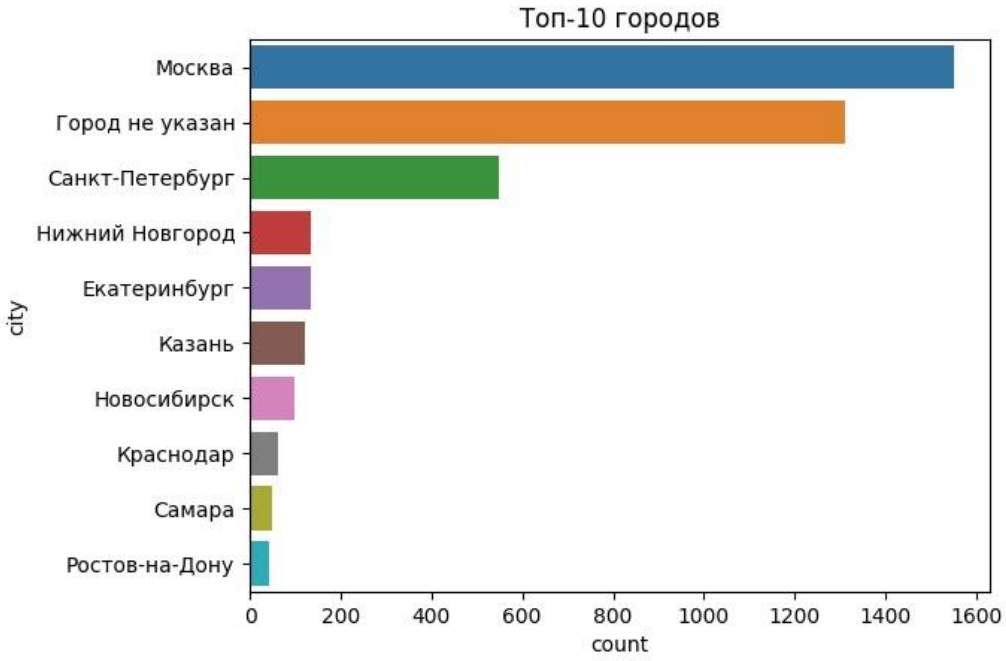


Вывод: на рынке IT у более чем 60% вакансий есть возможность работать удаленно.

3.2 Анализ предложений по вакансиям в зависимости от города

Безусловным лидером по количеству предложений – это город Москва с 1552 вакансиями, на втором месте идет Санкт-Петербург 548 вакансий. Среди региональных городов лидируют Нижний Новгород, Екатеринбург и Казань, у них примерно по 130

предложений, а это в 10 раз меньше чем в Москва. Отсюда можно сделать вывод, что шансы найти работу в IT сфере в Москве в 10 раз, а порой и в 20 раз больше чем в других городах.



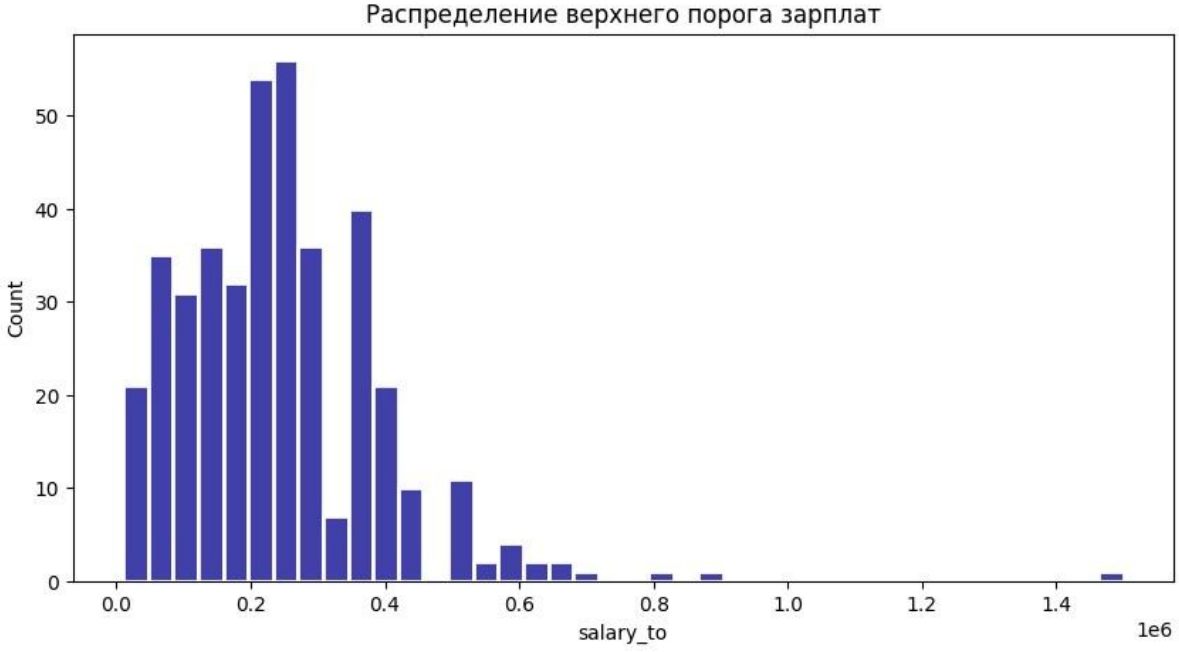
	city	count
0	Москва	1552
1	Город не указан	1310
2	Санкт-Петербург	548
3	Нижний Новгород	135
4	Екатеринбург	133
5	Казань	120
6	Новосибирск	98
7	Краснодар	61
8	Самара	50
9	Ростов-на-Дону	43

3.3 Анализ уровня заработной платы

Уровень заработной платы отражен только в 18% вакансий, поэтому выводы и оценки, сделанные на основании этой небольшой выборки, носят приблизительный характер, но в целом отражают действительную картину, которая складывается в настоящий момент на рынке труда.

- Среднее значение по нижней границе (salary_from) 158 000 рублей
- Среднее значение по верхней границе (salary_to) 241 000 рублей
- Стандартные отклонения в 105 000 (salary_from) и 152 000 (salary_to) говорят нам о том, что значения зарплат очень сильно разбросано вокруг своего среднего как в большую, так и меньшую сторону.
- Минимальный уровень заработной платы это 5000 рублей
- Максимальный уровень заработной платы это 1.5 миллиона рублей
- 50 % всех предложений в нижней границе (salary_from) находятся между 80 000 и 200 000 рублей
- 50% всех предложений в верхней границе (salary_to) находятся между 150 000 и 316 000 рублей

	salary_from	salary_to
count	532.0	404.0
mean	158760.0	241037.0
std	105585.0	152840.0
min	5000.0	10000.0
25%	80000.0	150000.0
50%	150000.0	220000.0
75%	200000.0	316250.0
max	800000.0	1500000.0



Наиболее часто встречающемся значением зарплаты в нижнем диапазоне является 200 000 рублей. Такое зарплату чаще всего предлагают в Сбере.

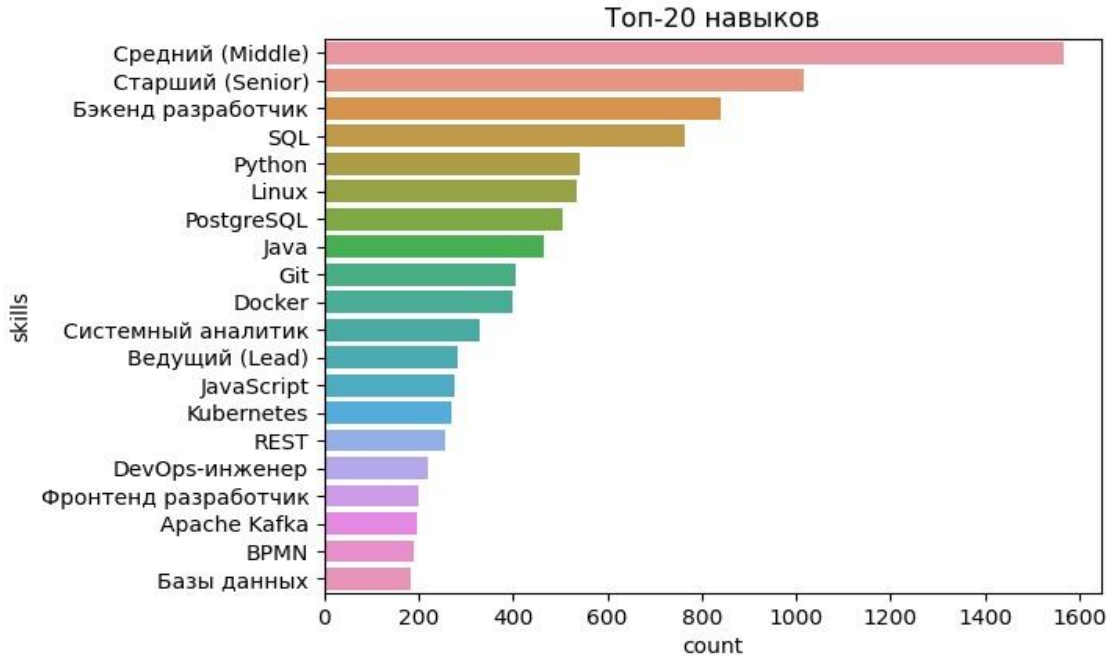
200000.0	65
150000.0	50
250000.0	39
100000.0	38
120000.0	31
180000.0	23
60000.0	20
80000.0	18
300000.0	18
50000.0	18

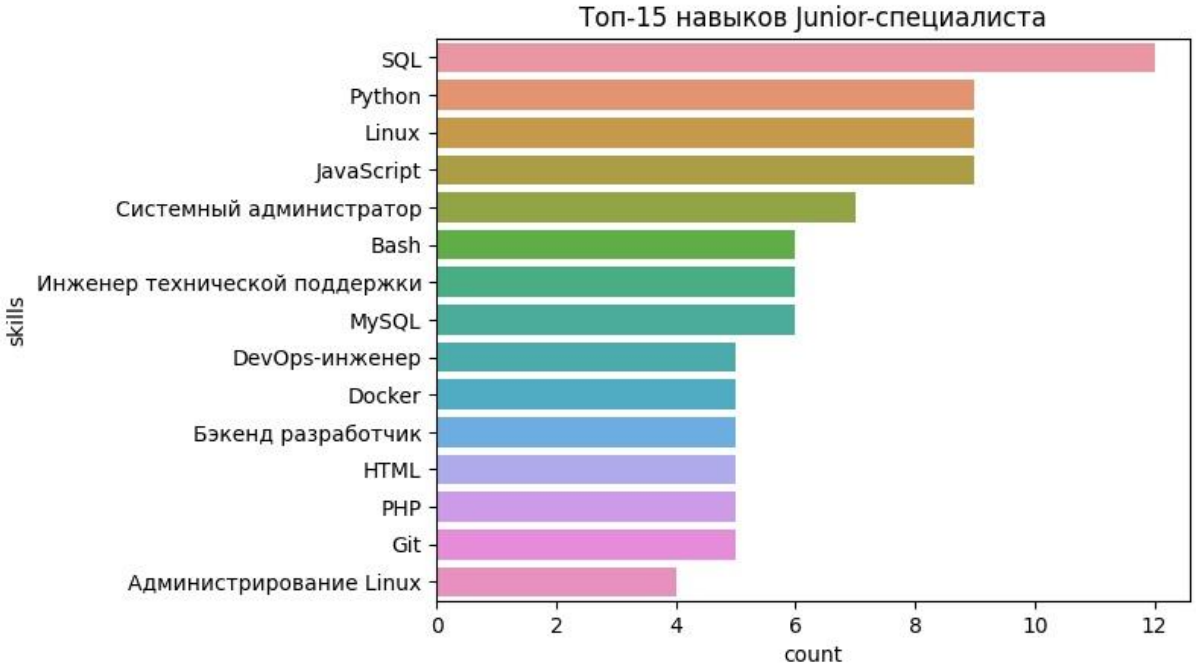
company	
Сбер	16
МегаФон	3
Heaad	2
PROMO IT	2
Pushflow	2
Black Shark Recruiting	2
Stivisto Inc.	1
Бристоль	1
Вебмониторэкс	1
Karma8	1

Вывод: 50% предложений на рынке по нижней границе лежат в диапазоне от 80 000 до 200 000 рублей, в верней же границе от 150 000 до 316 000 рублей.

3.4 Анализ наиболее востребованных навыков (hard skills)

- В настоящий момент на рынке труда наиболее востребованы специалисты уровня middle и senior, в 74% размещенных вакансиях ищут специалистов именно с таким опытом.
- Среди профессий лидирует backend-разработчик с 22%.
- Наиболее востребованным навыком является SQL 33%, на втором месте Python 15%, на третьем Linux 14%, Java 13%.
- Востребованность junior-специалистах крайне низкая, всего 2%
- Навыки, которыми должны обладать сотрудники, претендующие на позицию junior, напрямую коррелируются с наиболее востребованными навыками на рынке. Это SQL, Python, Linux.





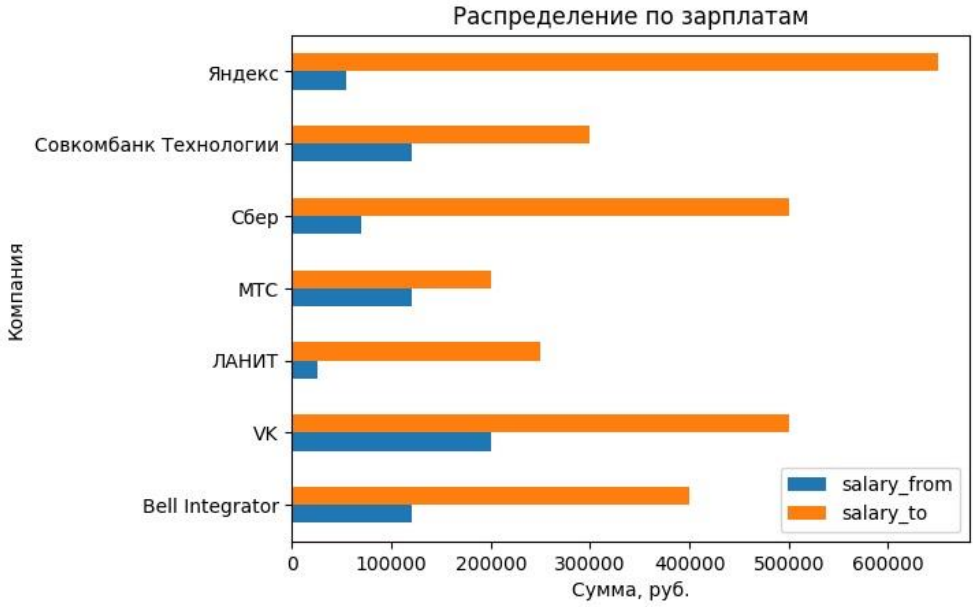
3.5 Анализ компаний по количеству размещенных вакансий

Сделаем список из топ-20 компаний по количеству размещенных вакансий. Лидером является Bell Integrator, более 300-та вакансий, компания предоставляет большой спектр IT-решений для бизнеса. В пятерку лидеров входят крупные российские финтех-компании: МТС, Сбер, Тинькофф, Банк ПСБ, VK.
Наиболее часто встречающейся вакансией в этом списке (топ-20), является «Системный аналитик» более 35 предложений и «Java-разработчик» 33 предложения.


```
[ ]  df_top_20['vacancy_name'].value_counts()

Системный аналитик                          35
Разработчик Java                            17
Java разработчик                            16
DevOps инженер                              13
Data engineer                               12
                                            ..
Android developer (МТС eСпортс)              1
Специалист технической поддержки             1
Team Lead / Senior PHP разработчик           1
Системный аналитик (Серверная Астра)         1
Инженер эксплуатации средств защиты сети     1
Name: vacancy_name, Length: 1539, dtype: int64
```

Самок крупное предложение по заработной плате, по нижней границе, предлагает VK – 200 000 рублей, а вот по верхней границе больше всего платит Яндекс – более 600 000 рублей готовы отдать за хорошего специалиста. А МТС, например, показывает диапазон средних значений по рынку.



Итоги проекта и заключение

По бизнес-задачам:

- В ходе исследования было выявлено, что в настоящий момент рынок в первую очередь нуждается в специалистах уровня middle и senior. Более чем в 60% случаев работодатель готов предложить сотруднику удаленный формат работы.
- Наиболее востребованной профессией на рынке является backend-разработчик.
- Работодатели в первую очередь хотят, чтобы сотрудники владели следующими навыками: SQL, Python, Linux, Java.
- Самой востребованный профессией является «Системный аналитик».
- В среднем работодатели готовы платить сотруднику от 80 000 до 316 000 рублей, все конечно сильно зависит от опыта и навыков.
- Самым привлекательным городом для поиска работы является город Москва.

Рекомендации стейкхолдерам:

Для образовательных порталов (IT-курсы, школы, университеты) – подбирать и корректировать учебные программы исходя востребованности на рынке специалистов и специальностей. При проведении маркетинговых компаний использовать актуальные и достоверные данные, отражающие текущую ситуацию. Определять компании, которые наиболее остро нуждаются в сотрудниках для проведения с ними партнерских программ, и подготовки необходимых специалистов.

Для работодателей (IT-компании) – оценивать ситуацию на рынке труда, по уровню заработной платы, своевременно индексировать и корректировать, чтобы постоянно находиться в рынке и избегать оттока сотрудников к конкурентам, если причина только в финансовой части. Информация о городах, так же позволяет находить свободные позиции для открытия филиалов и оценки действий конкурентов.
Т.к. более 60 процентов вакансий предусматривают удаленный формат работы – это дает отличную возможность развивать свою деятельность в регионах, позволяя тем самым существенно снижать издержки на аренду офисов и помещений.