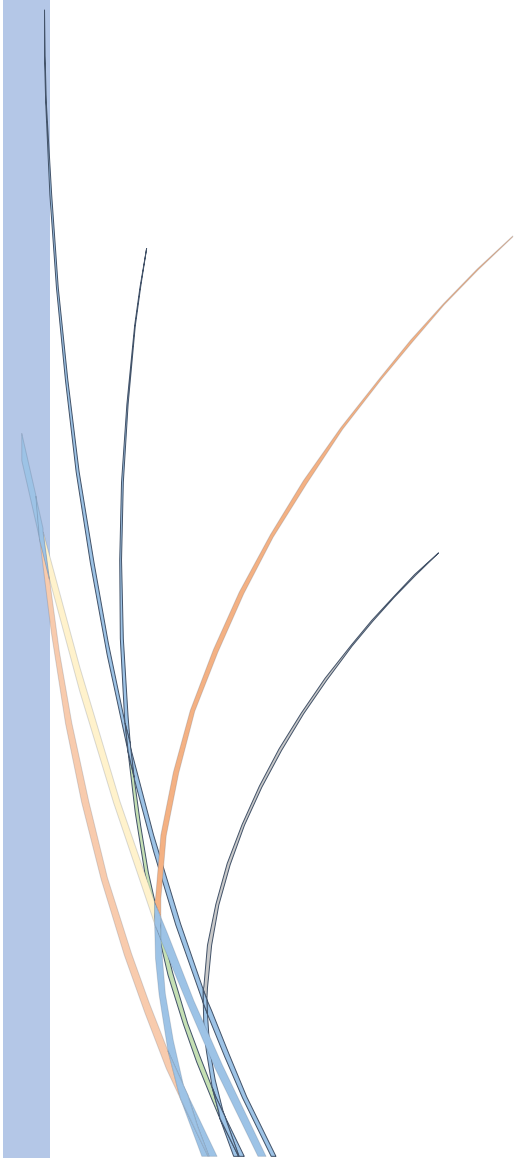
23.3.2024

Тестовое задание

Тема: Рынок труда



Вишталюк Г.С.

Оглавление

Проблема

Выпусник цифровых кафедр ТГУ по специальности «Аналитик данных», обладает необходимыми знаниями и практическими навыками в работе с данными (сбор, обработка, аналитика), BI-системами, базами данных (конструирование, знание SQL/NoSQL баз, обращение к ним). Молодой человек готов к релокации в границах Российской Федерации, либо может остаться в Томске. Какие города России будут наиболее привлекательными по заработной плате ?

Цель

Используя доступные источники данных определить наиболее привлекательные по заработной плате для релокации города России

Источники данных

В рамках данной работы использовались актуальные на 22 марта 2024 года открытые вакансии с сайта HH.ru

Поиск открытых датасетов для выполнения проекта

Для выполнения проекта был написан скрипт-парсер на языке Python. Парсер разработан для работы с API сайта HH.ru. Осуществлялся поиск вакансий по запросам: BI, SQL, Data, Аналитик (список запросов составлялся исходя из профессиональных навыков соискателя).

Вакансии, предоставляемые HH.ru по запросам, проходили следующую фильтрацию:

- Содержание в названии вакансии ключевых слов: bi, sql, данные, data, аналитик, analyst
- Указана зарплата (нижняя граница, верхняя граница, зарплатная вилка)

По результатам работы парсера был получен файл в формате csv. Содержащий в себе следующие столбцы:

- Name: название вакансии
- Min_salary: нижняя граница зарплатной вилки
- Max_salary: верхняя граница зарплатной вилки
- Currancy: валюта, в которой указаны заработная плата
- Location: город

Данный датасет так же был подвержен дальнейшей фильтрации: были удалены записи о вакансиях, которые не соответствовали запросу, но прошли автоматическую фильтрацию, так же были отсеяны записи о вакансиях, расположенных не в городах России.

Таким образом был получен итоговый датасет с 1894 записями информации об актуальных вакансиях с сайта HH.ru

Анализ

Итоговый датасет для анализа содержит 1894 записи вакансий. Интерес представляют вакансии только внутри России и указанием зарплатной вилки. К сожалению, не все работодатели указывают полную зарплатную вилку, поэтому для анализа так же были взяты вакансии с указанием нижней, либо верхней границы заработной платы. Так же не все записи в качестве основной валюты использовали Российский рубль (RUR), в датасете имеются записи, в которых заработная плата выражена в долларах (USD), либо евро (EUR)

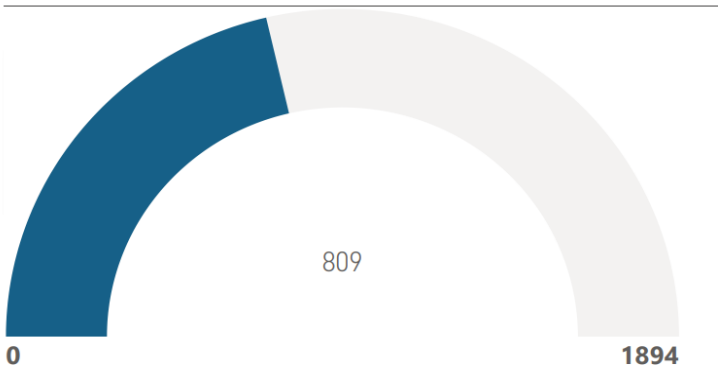
Этап 1

Поэтому на первом этапе исследования все вакансии были приведены к единой валюте. Для этого были выбраны все вакансии в иностранной валюте, и их верхняя и нижняя зарплатные границы были пересчитаны в рубль (RUR) по следующему курсу по отношению к рублю (на 19.03.2024):

- Евро (EUR): 99,22
- Доллар (USD): 91,45

Этап 2

Следующий этап анализа заключался в заполнении пропущенных значений вилок:



Количество вакансий с полной зарплатной вилкой в итоговом датасете составило 809 записей или 42,7% (Рис.1). Для каждой полной записи были рассчитаны зарплатные интервалы, как разница между границами вилки.

Рис. 1 Количество вакансий с полной зарплатной вилкой

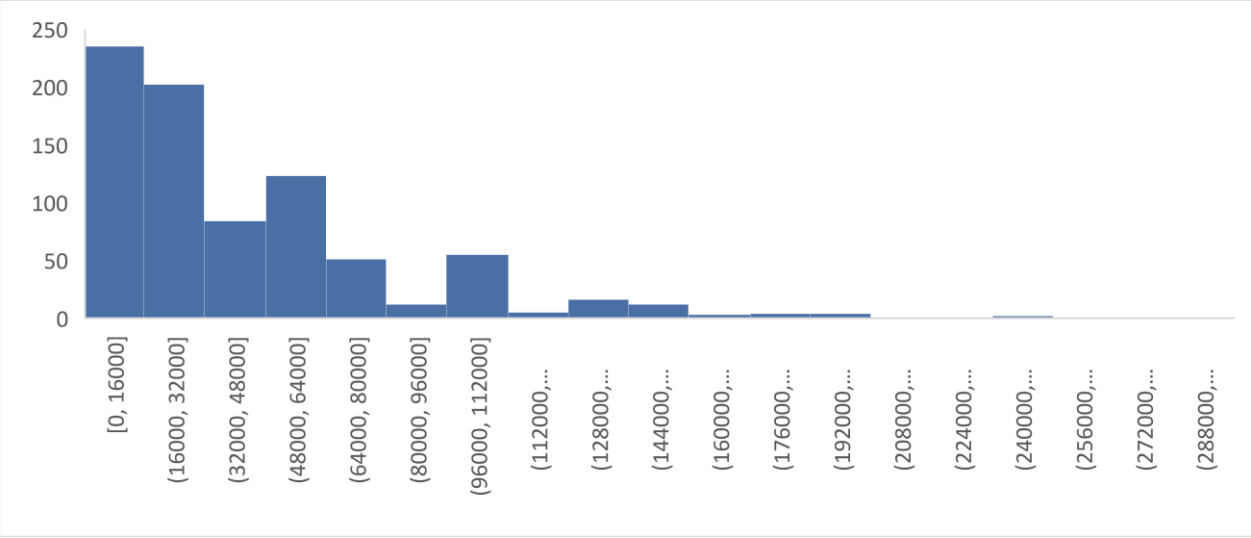


Рис. 2 Распределение значений интервалов зарплатной вилки

Для полученных интервалов зарплатной вилки была построена гистограмма распределения, из внешнего вида которой следует, что распределение данной величины не соответствует нормальному, значит в качестве основной характеристики была выбрана медиана.

Медиана значений интервалов зарплатной вилки составила 30000 руб.

Таким образом, на основе полученного значения, в основном датасете были заполнены недостающие значения зарплатных вилок следующим образом:

$$\text{нижняя граница} = \text{верхняя граница} - 30000$$
$$\text{верхняя граница} = \text{нижняя граница} + 30000$$

В случае, если вычисленная нижняя граница была меньше нуля, это значение приравнивается нулю

Также на данном этапе для каждой записи были выведены средние значения заработной платы – середина интервала зарплатной вилки

Эти данные отображены в дополнительном столбце Avg_salary

Этап 3

На третьем этапе аналитической работы производилась оценка распределения вакансий среди городов России.

В каждой записи о вакансии имеется информация о локализации:

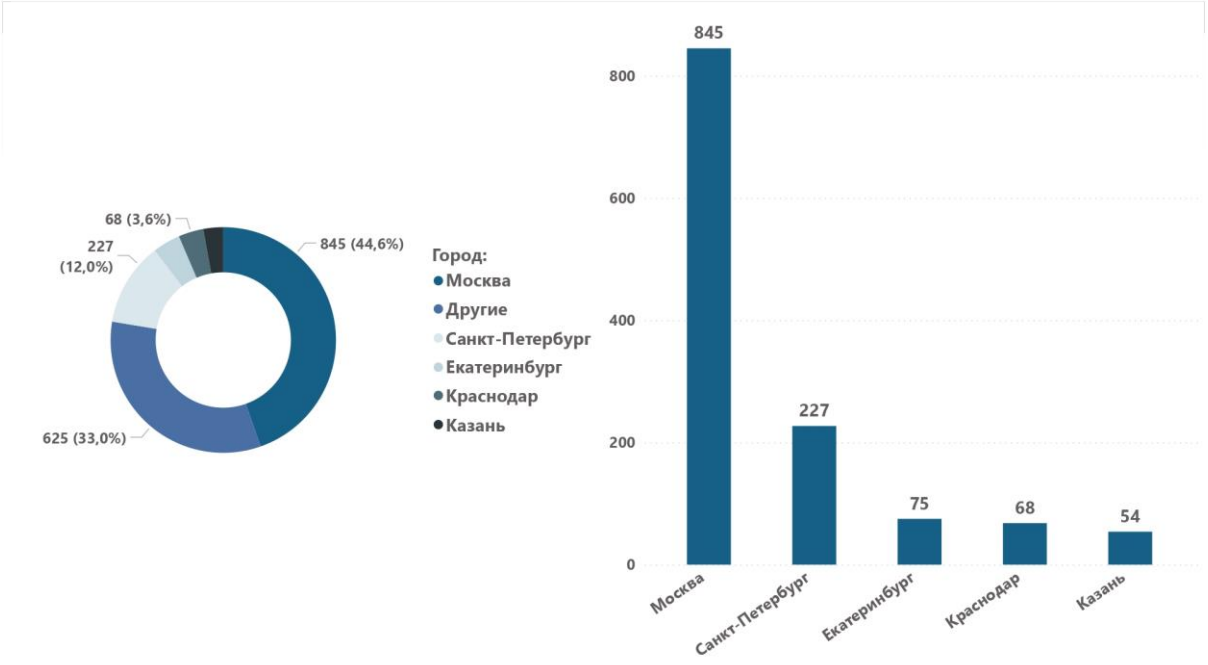


Рис 3. Распределение вакансий по городам.

Таким образом был составлен список наиболее привлекательных городов с точки зрения количества вакансий для соискателя.

Топ-5:

1. Москва – 845 вакансий (44,6%)
2. Санкт-Петербург – 227 вакансий (12%)
3. Екатеринбург – 75 вакансий (3,9%)
4. Краснодар – 68 вакансий (3,6%)
5. Казань – 54 вакансии (2,9%)

В процессе работы, все вакансии были распределены по 3-м соответствующим направлениям. Доли среди них распределены следующим образом:

1. Аналитик данных – 89,2 %
2. Базы данных – 7 %
3. BI-аналитик – 3,7%

Также во внимание была принята локализаций вакансий, их распределение по городам РФ, с целью определения главных мест сосредоточения всех специальностей.

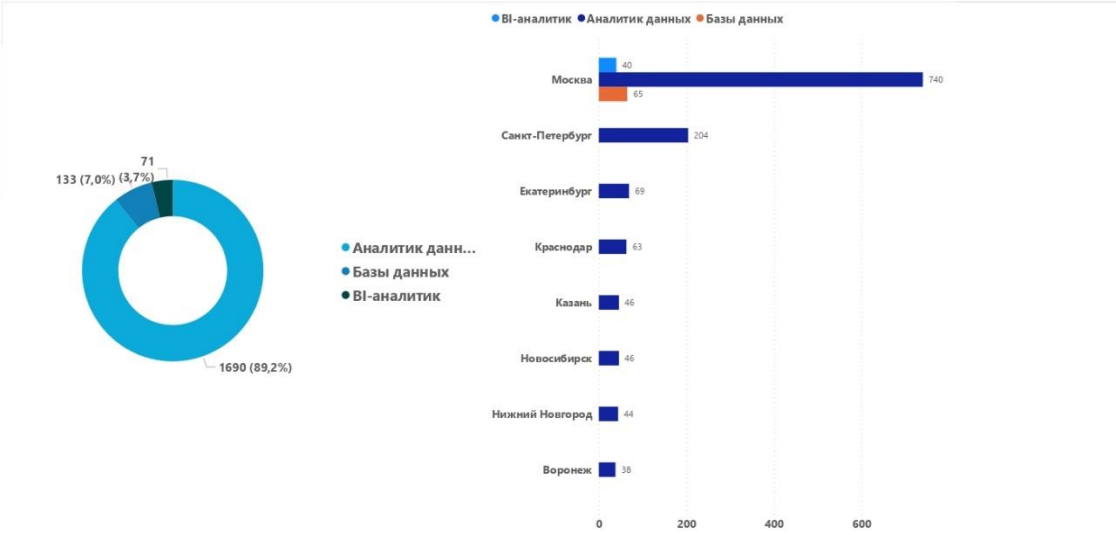


Рис. 4 Распределение направлений

Для каждого направления были составлены списки городов с наибольшим количеством вакансий.

Таким образом были составлены топ-3 города для каждого из направлений:

Лидирующие позиции занимают Москва (740 вакансий для аналитиков, 40 – для BI-аналитиков, 65 для специалистов по работе с базами данных) и Санкт-Петербург (204 вакансий для аналитиков, 8 – для BI-аналитиков, 15 для специалистов по работе с базами данных), а на 3-ем месте у каждого из направлений разместились свои города:

- Аналитика данных: Екатеринбург – 69 вакансий
- BI-аналитик: Самара – 3 вакансии
- Базы данных: Казань – 7 вакансий

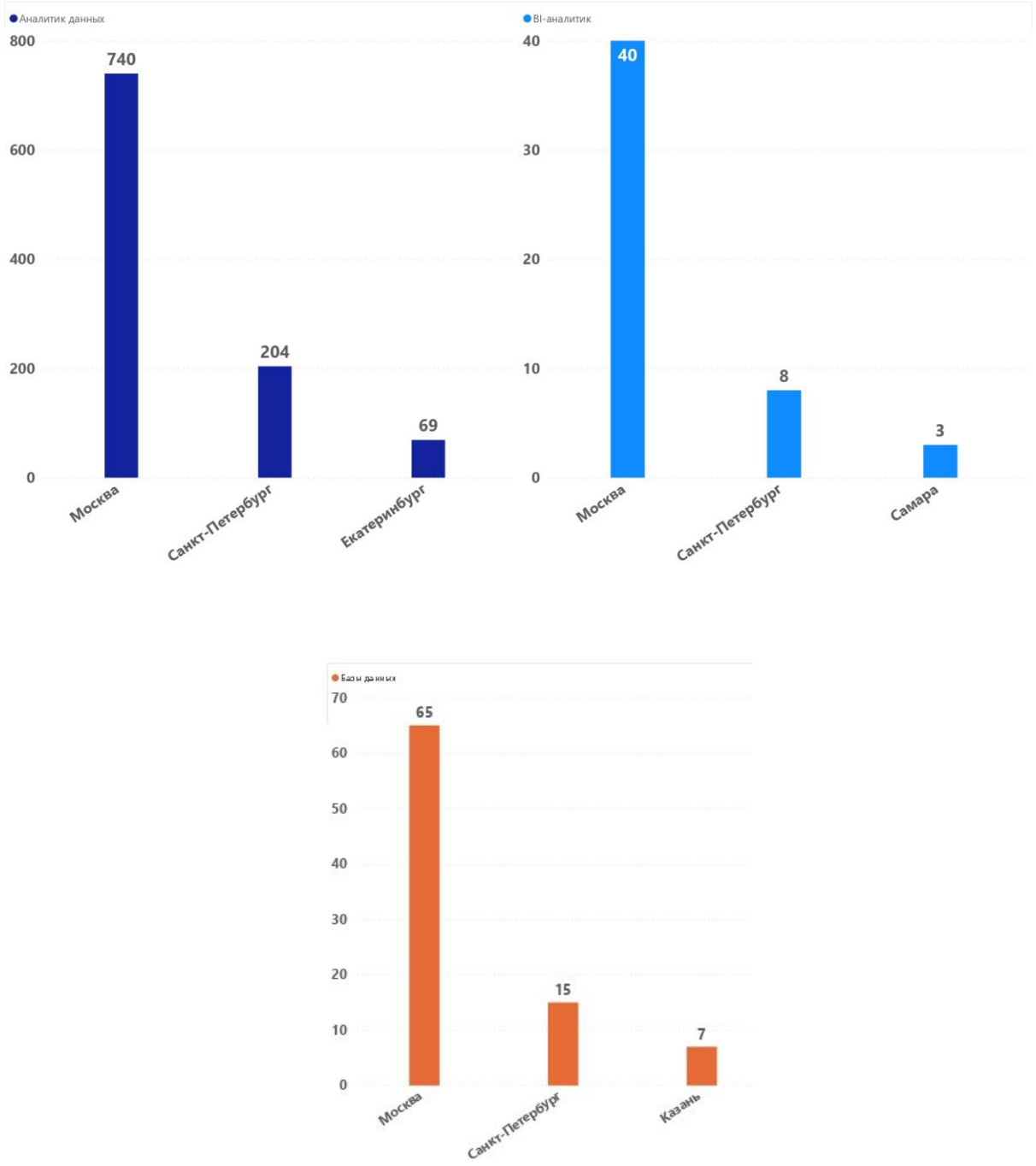


Рис. 5 Топ-3 города по направлениям

Этап 4

На четвертом аналитическом этапе рассмотрены уровни заработной платы в различных городах. Было рассмотрено три критерия сравнения: нижняя граница зарплатной вилки, верхняя граница и середина интервала.

Список городов проходил фильтрацию по количеству вакансий. Для анализа брались:

- Аналитик данных: больше либо равно пяти вакансий на город
- BI-аналитик и Базы данных: больше либо равно двух вакансий на город

В качестве описательной статистики для оценки уровня заработной платы была выбрана медиана, т.к. после проведённой оценки распределения данных критериев сделан вывод о несоответствии нормальному закону распределения.

Аналитик данных

По результатам по направлению Аналитик данных были составлен топ-5 городов, наиболее привлекательных по уровню заработной платы:

Нижняя граница:

1. Москва – 130000 руб.
2. Санкт-Петербург – 100000 руб.
3. Калининград – 90000 руб.
4. Ярославль – 90000 руб.
5. Владивосток – 80000 руб.

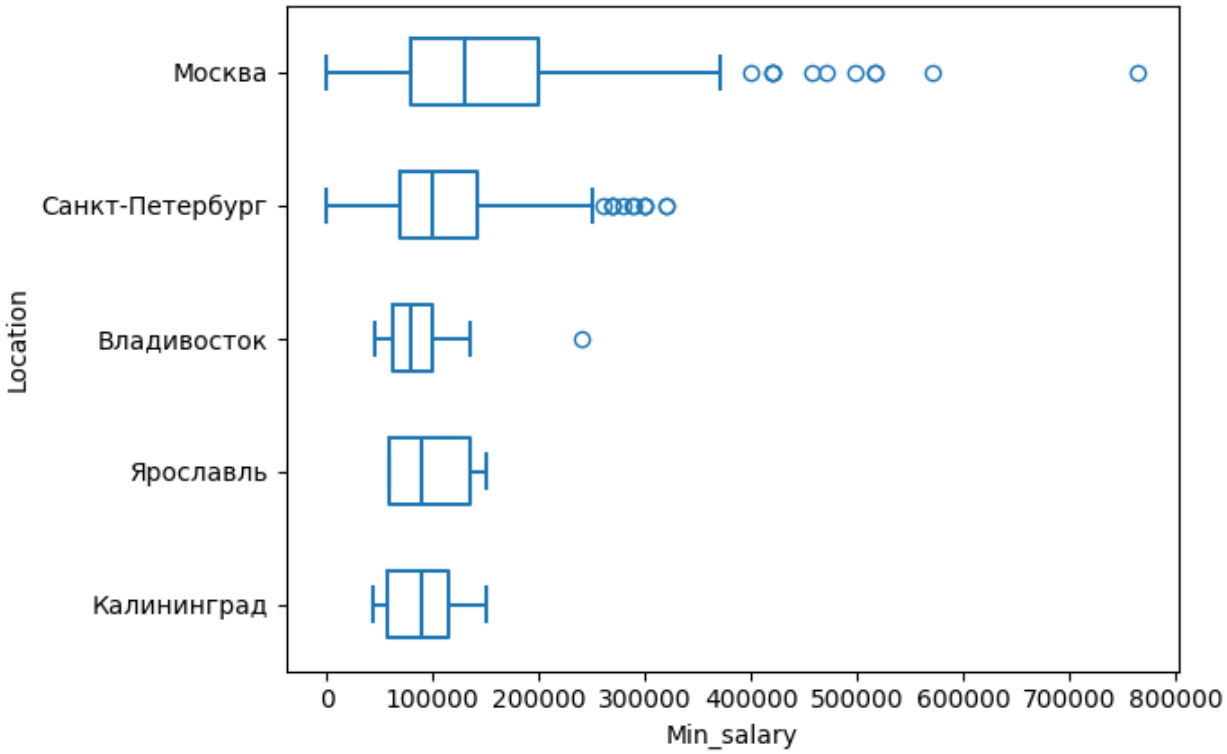


Рис. 6 Нижняя граница по направлению «Аналитик данных»

Верхняя граница:

1. Москва – 170000 руб.
2. Калининград – 147500 руб.
3. Санкт-Петербург – 100000 руб.
4. Ярославль – 130000 руб.
5. Набережные Челны – 104000 руб.

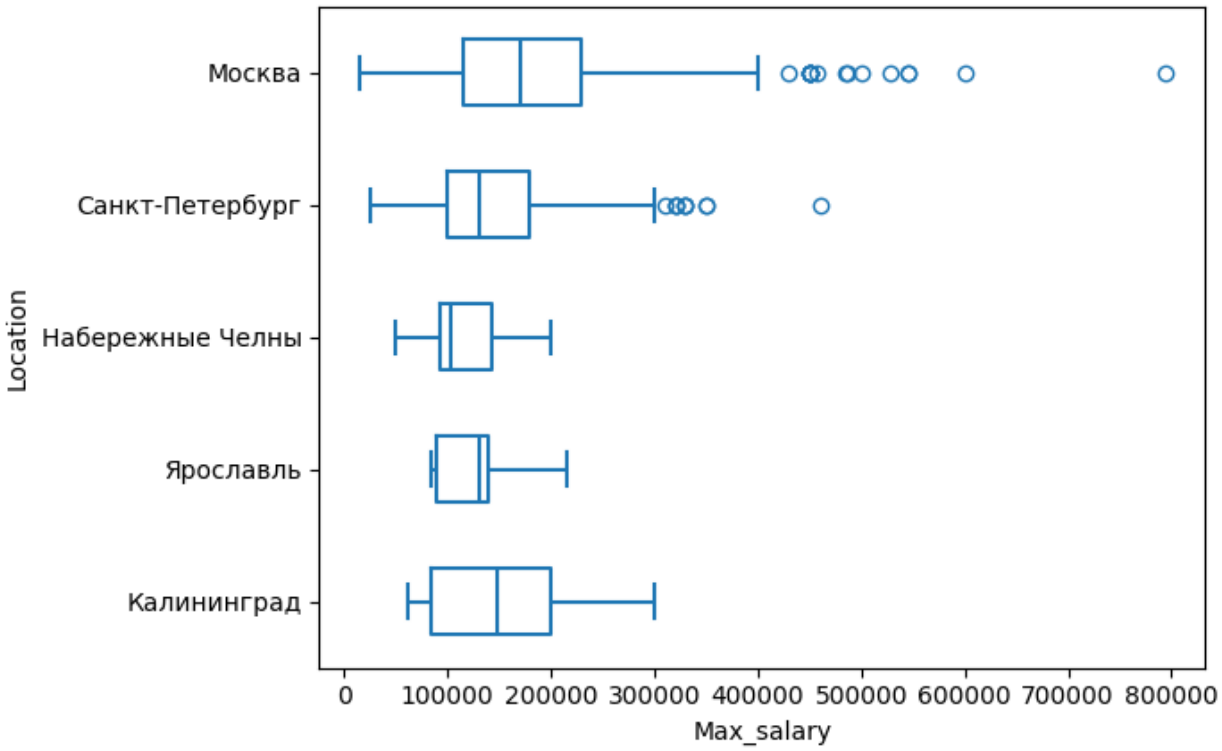


Рис. 7 Верхняя граница по направлению «Аналитик данных»

Средняя заработная плата:

1. Москва – 150000 руб.
2. Калининград – 118750 руб.
3. Санкт-Петербург – 115000 руб.
4. Ярославль – 115000 руб.
5. Набережные Челны – 89000 руб.

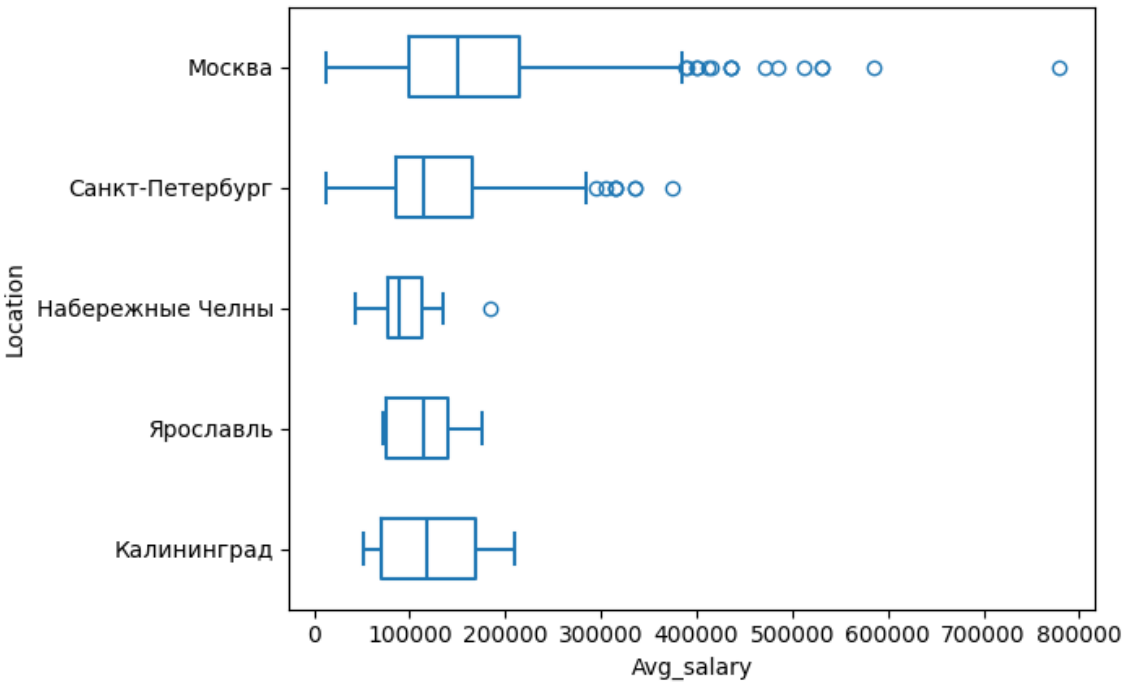


Рис. 7 Средняя заработная плата по направлению «Аналитик данных»

BI-аналитик

Для направления BI-аналитик топ-5 городов не зависел от критерия оценки, по всем категориям заработных плат этот список оставался неизменным. Города расположены следующим образом:

Место	Город	Нижняя граница, руб.	Средний уровень, руб.	Верхняя граница, руб.
1	Москва	194200	207950	227500
2	Санкт-Петербург	150000	165000	200000
3	Екатеринбург	133000	140500	148000
4	Новосибирск	85000	102500	120000
5	Самара	70000	75000	90000

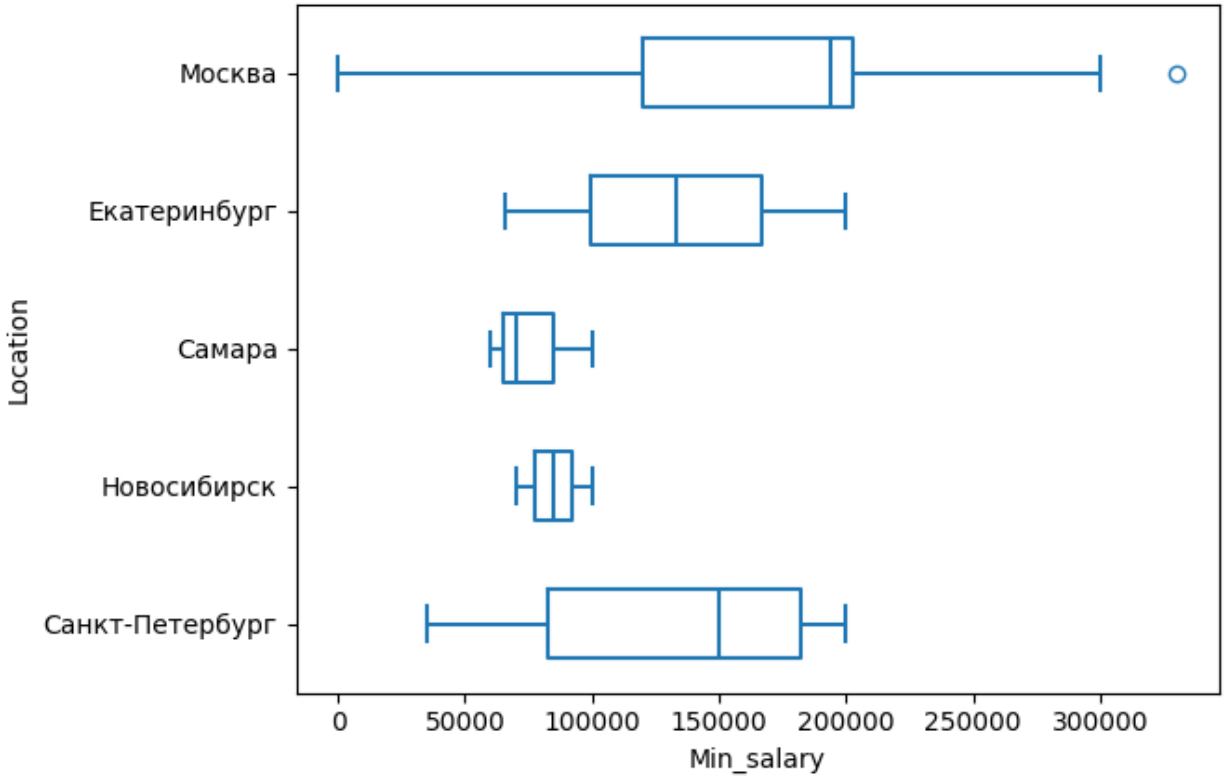


Рис. 8 Нижняя граница по направлению «BI-аналитик»

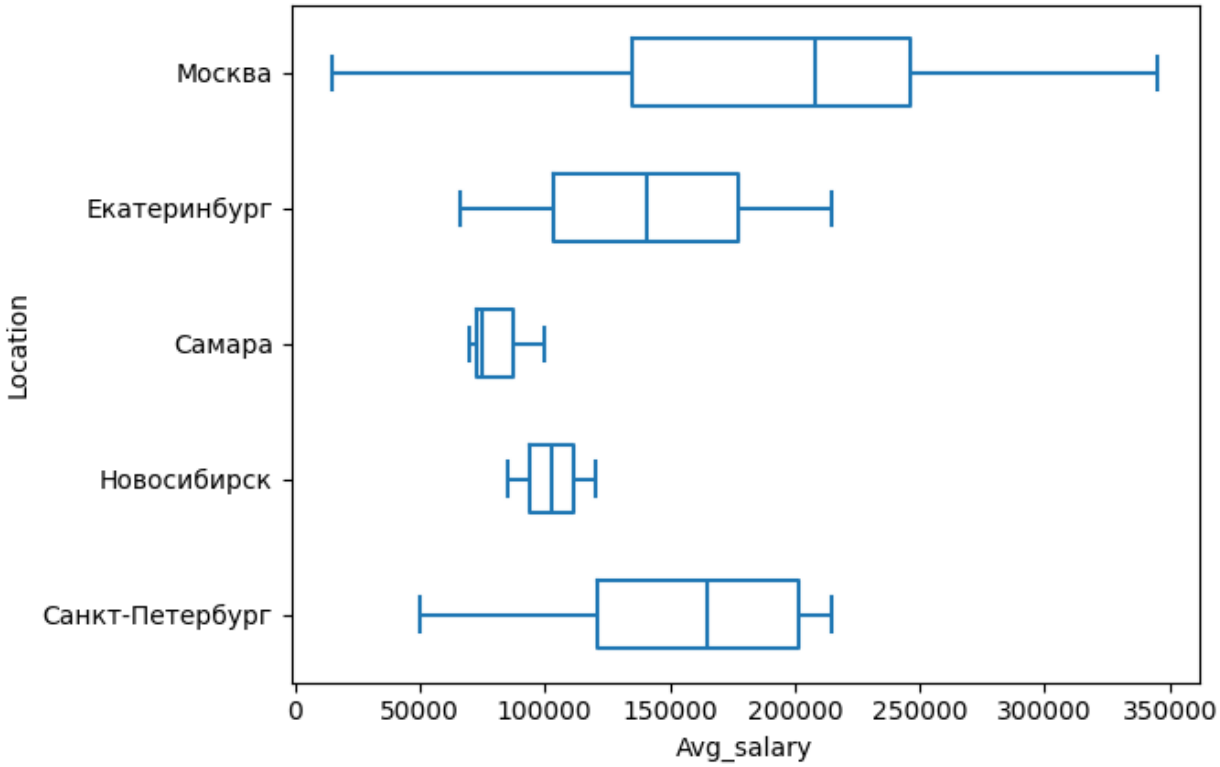


Рис. 9 Средний уровень по направлению «BI-аналитик»

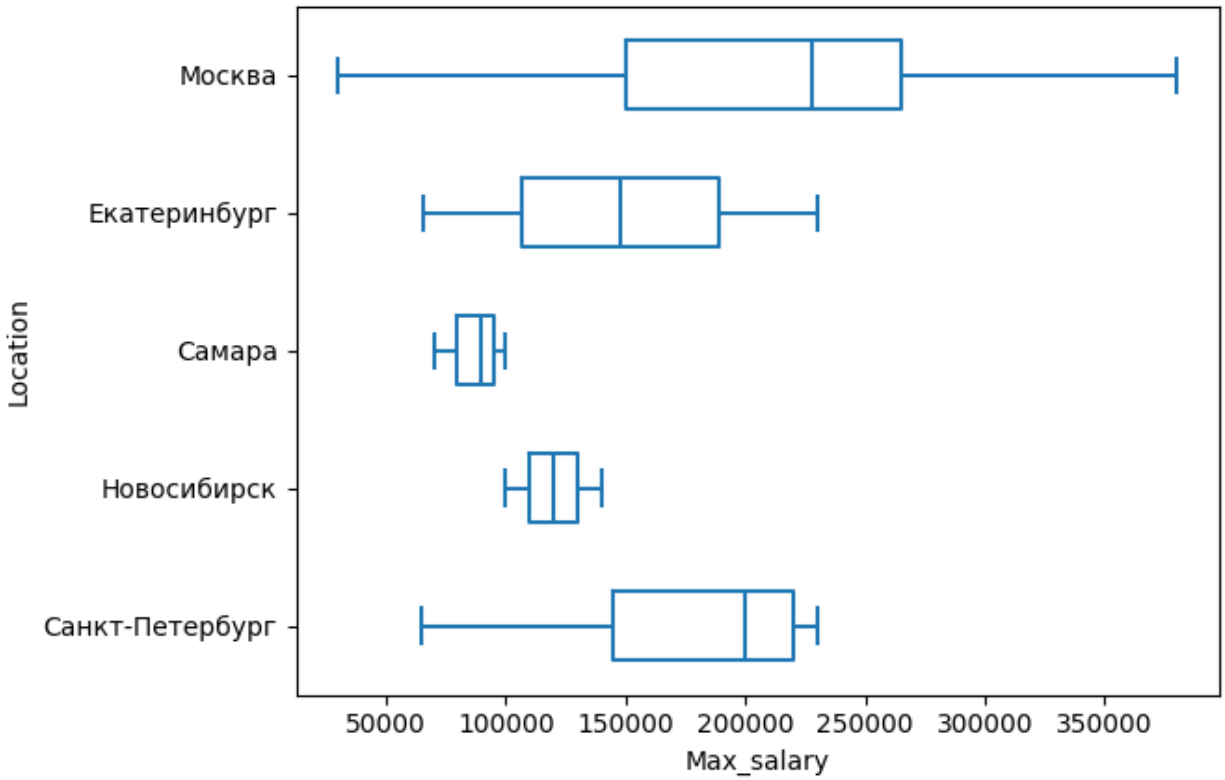


Рис. 10 Верхняя граница по направлению «BI-аналитик»

Базы данных

Нижняя граница:

1. Москва – 170000 руб.
2. Ростов-на-Дону – 170000 руб.
3. Краснодар – 165000 руб.
4. Екатеринбург – 150000 руб.
5. Санкт-Петербург – 150000 руб.

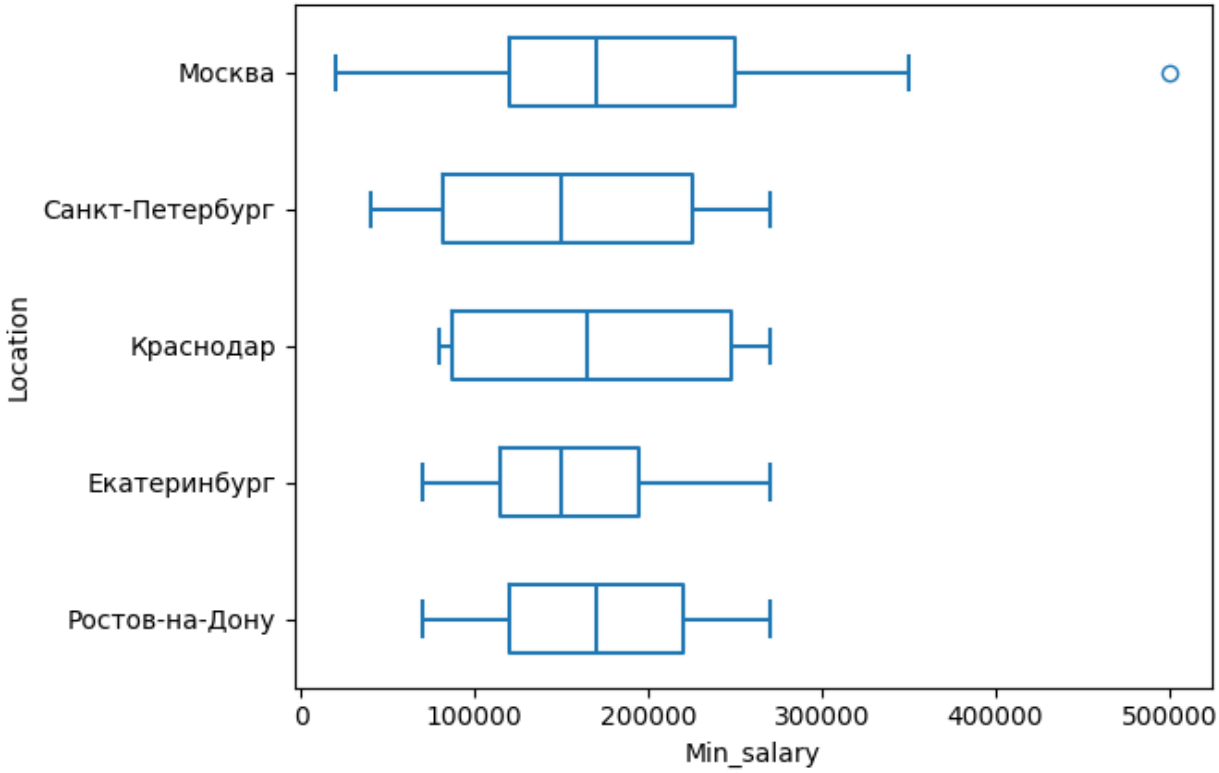


Рис. 11 Нижняя граница по направлению «Базы данных»

Верхняя граница:

1. Москва – 230000 руб.
2. Владивосток – 218500 руб.
3. Ростов-на-Дону – 215000 руб.
4. Краснодар – 210000 руб.
5. Екатеринбург – 205000 руб.

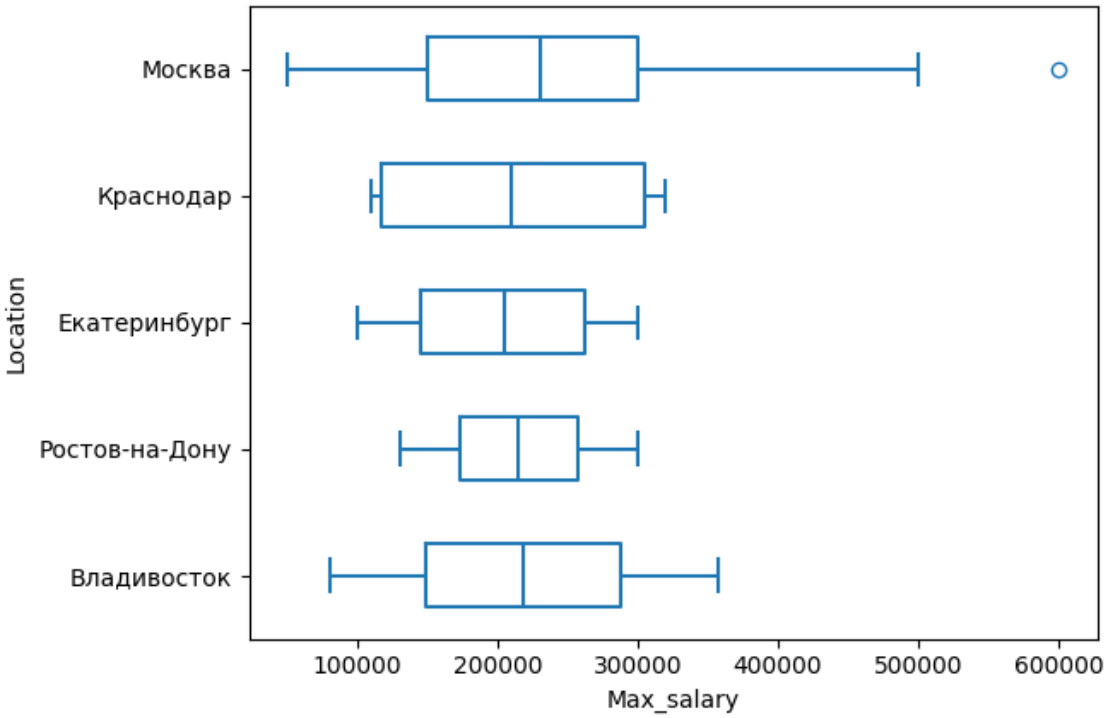


Рис. 12 Верхняя граница по направлению «Базы данных»

Средний уровень:

1. Краснодар – 192500 руб.
2. Ростов-на-Дону – 192500 руб.
3. Москва – 185000 руб.
4. Екатеринбург – 177500 руб.
5. Владивосток – 174000 руб.

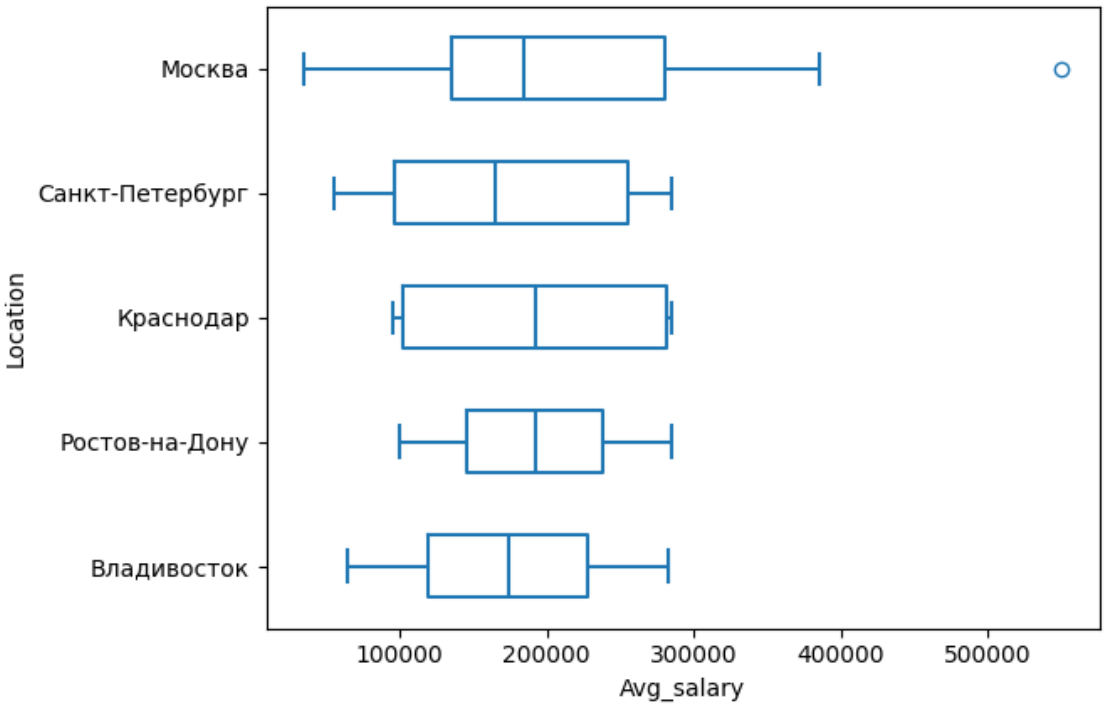


Рис. 13 Средний уровень по направлению «Базы данных»

Заключение

Таким образом, исходя из проведённого анализа, можно сделать вывод о том, что, все три направления сильно отличаются друг от друга, но можно определить сделать некоторые обобщённые вывод:

1. Исходя из анализа распределения направлений, наиболее часто попадаются вакансии «Аналитик данных», соответственно потенциальному соискателю будет проще искать место именно в этом направлении
2. Среди проанализированных вакансий чаще всего встречаются крупнейшие города России: Москва и Санкт-Петербург. Соискателю следует обратить внимание именно на них.
3. Анализ заработной платы по каждому направлению показал, что наиболее привлекательными по уровню заработка городами являются: Москва, Санкт-Петербург, Екатеринбург.