Федеральное государственное автономное образовательное учреждение

высшего образования

«Санкт-Петербургский политехнический университет Петра Великого»

Институт компьютерных наук и технологий

Высшая школа программной инженерии

АНАЛИЗ РЫНКА ВАКАНСИЙ IT-СПЕЦИАЛЬНОСТЕЙ С HEADHUNTER

**Курсовой проект**

по учебной дисциплине

«Наука о данных и аналитика больших объемов информации»

Направление подготовки магистров 01.04.02 «Прикладная математика и информатика»

Группа 5040102/30201

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнили студенты: |  |  | **Завьялов В. В.**  **Завьялов И. В.** |
|  |  |  |  |
| Руководитель: |  |  | **Никифоров И. В** |

Санкт-Петербург

2023

# Введение

Когда мы говорим об анализе рынка вакансий, мы исследуем различные аспекты рынка труда в IT-сфере: от востребованных навыков и специализаций до уровня заработной платы и месторасположения вакансий. HeadHunter, как крупнейшая российская платформа для поиска работы и персонала, предоставляет обширную базу данных вакансий, которую можно проанализировать.

Используя данные HeadHunter, мы можем оценить текущее состояние рынка IT-вакансий, выявить тенденции в требованиях к кандидатам, определить востребованные навыки и технологии, а также проанализировать региональное распределение вакансий. Этот анализ поможет компаниям, кандидатам и отрасли в целом понять динамику рынка труда, что важно для принятия обоснованных карьерных и бизнес-решений.

Проведение анализа рынка вакансий IT-специальностей с использованием данных HeadHunter имеет большое практическое значение для широкого круга заинтересованных лиц.

# Обзор предметной области

В современном мире информационных технологий вакансии IT-специалистов имеют особое значение. Понимание текущих тенденций и востребованных профессий является ключевым для успешной карьеры в области информационных технологий. Компании и предприятия нуждаются в профессионалах, способных эффективно решать задачи в области разработки программного обеспечения, администрирования систем, анализа данных и многих других сферах. Однако перед нанимателями стоит сложная задача — найти подходящих кандидатов среди огромного количества соискателей.

Для того чтобы облегчить процесс подбора персонала и значительно сократить время, затрачиваемое на поиск, было разработано множество ресурсов и сервисов, предоставляющих информацию о вакансиях и соискателях.

Данное исследование будет базироваться на данных платформы HeadHunter, которая является одной из самых популярных и широко используемых для поиска работы и найма персонала в России и других странах СНГ.

HeadHunter — один из самых крупных сайтов по поиску работы и сотрудников в мире (по данным рейтинга Similarweb). Компания занимается тем, что создает передовые технологии на всех доступных платформах для того, чтобы работодатели могли быстро найти подходящего сотрудника, а соискатели — хорошую работу [1].

Для сбора данных мы использовали HeadHunter API. HeadHunter API предоставляет доступ к обширной базе данных вакансий, которая постоянно обновляется и увеличивается. Данный API очень удобный, имеет подробную документацию и привычный синтаксис запросов.

С помощью HeadHunter API вы можете получать информацию о соискателях и компаниях, а также использовать функциональность HeadHunter для своего сайта или приложения [2].

# Постановка задач анализа и актуальность

Анализ рынка вакансий IT-специальностей имеет большую актуальность в современном мире, где IT-отрасль считается одной из самых динамично развивающихся и востребованных. Результаты анализа позволят получить ценную информацию о состоянии рынка труда в IT-сфере. Это может быть полезно как для работодателей, которым необходимо понять текущие требования и тенденции в отношении найма IT-специалистов, так и для соискателей, которые могут использовать эти данные для определения наиболее востребованных навыков и специализаций.

Использование результатов анализа предоставляет возможность принимать обоснованные решения в отношении карьерного развития, обучения и планирования ресурсов.

Основная задача данного проекта состоит в анализе рынка вакансий IT-специальностей с использованием данных, предоставленных платформой HeadHunter.

Сформулированные нами аналитические задачи можно разделить на 5 подгрупп:

* Сравнительный анализ предлагаемых заработных плат в IT-сфере. Исследование уровня заработной платы в различных сегментах IT-рынка и сравнительный анализ оплаты труда в зависимости от города, профессиональной роли и типа занятости;
* Исследование распределения требуемого опыта от соискателей и анализ оплаты труда в зависимости от опыта;
* Определение топ-работодателей. Определение компаний, активно размещающих вакансии в IT-сфере, и анализ их потребностей в персонале, ключевых навыках и опыте работы;
* Определение наиболее востребованных навыков для профессии программист разработчик;
* Тепловая карта IT-вакансий.

# Данные

Для получения данных использовался API HH. API HH – это инструмент, позволяющий разработчикам взаимодействовать с базой данных вакансий и резюме, предоставленных на платформе HH.ru. API предоставляет доступ к обширной информации о вакансиях, работодателях, резюме и кандидатах, что делает его мощным инструментом для анализа рынка труда и поиска работы.

Некоторые возможности API HH:

* Получение списка вакансий. API HH предоставляет возможность получить список вакансий с различными параметрами поиска, такими как ключевые слова, местоположение, категория, зарплата и т.д. Это позволяет пользователям настроить поиск под свои требования и получить актуальную информацию о доступных вакансиях.
* Получение информации о вакансии. API позволяет получить детальную информацию о конкретной вакансии, включая описание, требования, компанию-работодателя и т.д.

Ограничения API HH:

* Ограничение на количество запросов: API HH имеет ограничение на количество запросов, которые можно выполнить в определенном временном интервале. Как правило, есть ограничение на количество запросов в минуту или в день. При превышении лимита API может вернуть ошибку или ограничить доступ к данным. Данное ограничение составляет 10 запросов/сек/ip. При превышении выдаётся ответ с HTTP-кодом 429 (too many requests)
* Ограничение на количество возвращаемых результатов в одном запросе. При выполнении запроса можно получить не более 2000 вакансий.
* Можно выкачать вакансии только за последние 30 дней

Для получения списка вакансий необходимо отправить GET-запрос на адрес «https://api.hh.ru/vacancies» [3]. Список вакансий фильтруется согласно переданным параметрам запроса. Параметры запроса, которые были использованы в работе, представлены в таблице 1. Пример ответа представлен на рисунке 1. В ответе можно увидеть id вакансии, ее название, заработная плата, специальность и т.д.

Таблица 1 – Параметры запроса.

|  |  |
| --- | --- |
| Параметр | Описание |
| page | Номер страницы |
| per\_page | Количество вакансий на страницу. |
| professional\_role | Профессиональная область. Необходимо передавать id роли. |
| date\_from | Дата, которая ограничивает снизу диапазон дат публикации вакансий. Значение указывается в формате ISO 8601 - YYYY-MM-DD или с точность до секунды YYYY-MM-DDThh:mm:ss±hhmm. |
| date\_to | Дата, которая ограничивает сверху диапазон дат публикации вакансий. Значение указывается в формате ISO 8601 - YYYY-MM-DD или с точность до секунды YYYY-MM-DDThh:mm:ss±hhmm. |

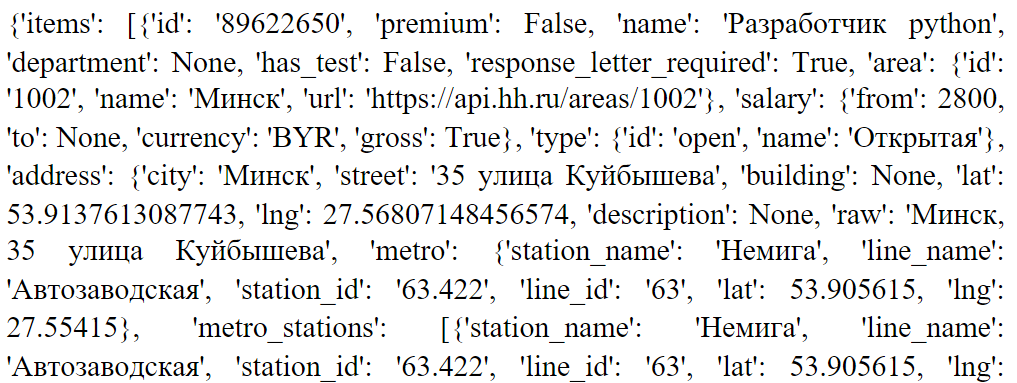


Рисунок 1 – Пример ответа.

Используемые запросы:

* GET /vacancies - возвращает список вакансий
* GET /vacancies/{vacancy\_id} - возвращает JSON с телом вакансии
* GET /professional\_roles - возвращает список категорий и ролей

В работе собирались только вакансии, которые относятся к сфере информационных технологий. Для этого необходимо было собрать id всех ролей из этой сферы. Чтобы это сделать, необходимо использовать GET запрос /professional\_roles, который возвращает список категорий. В этом списке находим категорию, которая имеет название “Информационные технологии". Из данной категории получаем список id ролей.

Чтобы получить тело каждой вакансии использовался GET запрос /vacancies/{vacancy\_id}, в котором указывался id определенной вакансии. Результат каждого запроса в формате JSON отправляется в базу данных.

Объём получившегося набора данных составил 2,1 Гб. Для сбора данных потребовалось 2 дня.

Количество собранных вакансий – 190880. Количество переменных 59.

Формат экспорта – CSV

Данные хранятся в нереляционной СУБД MongoDB в виде JSON-документов.

На рисунке 2 представлено количество найденных за день вакансий, и сколько из них удалось скачать. Данные представлены для 29 дней.

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Рисунок 2 – Количество собранных вакансий за день

# Обзор аналогичных исследований

Для проведения анализа рынка вакансий IT-специальностей можно обратиться к ранее проведенным исследованиям, в которых также анализировали рынок труда в IT-сфере.

Из наиболее интересных можно выделить периодически публикуемые HeadHunter исследования IT-рынка вакансий. Например, отчеты за 2019 и 2023 год.

Из особенностей отчета за 2023 год можно выделить анализ динамики уровня конкуренции за рабочие места с 2021 по 2023 год, а также анализ уровня конкуренции по языкам программирования и основным IT-профессиям.

Отчет HeadHunter за 2019 год интересен тем, что предоставляет 15 портретов представителей IТ-специальностей. Каждый портрет содержит:

* Самые необходимые навыки в специальности;
* Анализ востребованности специальности с 2016 по 2018 год;
* Анализ требуемого опыта работы;
* Навыки, на которые растет спрос;
* Медианные зарплаты в зависимости от опыта работы;

В статье “Зарплаты и конкуренция на IT-рынке труда в России сегодня” проводится таже аналитика, что и в отчете HeadHunter за 2023 год.

Аналогичные исследования:

* Crawler for retrieving data from headhunter: https://github.com/kelsonauts
* Зарплаты и конкуренция на IT-рынке труда в России сегодня: https://habr.com/ru/companies/hh/articles/759292/
* Обзор ИТ-отрасли по итогам первого полугодия: какие зарплаты платят и насколько сложно найти работу: https://kirishi.hh.ru/article/31783
* ИТ: обзор рынка вакансий и топ-15 специальностей: https://kirishi.hh.ru/article/24562

В предложенных исследованиях есть результаты некоторых аналитических задач, которые нам требуется решить. Результаты данных задач приведены на рисунках 3 и 4.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Рисунок 3 – Медианные предполагаемые зарплаты в IT сфере в зависимости от опыта работы

Изображение выглядит как снимок экрана, текст, Красочность, диаграмма

Автоматически созданное описание

Рисунок 4 – Используемые языки в крупных компаниях

# ПРОГРАММНое решение

## Архитектура проекта

Архитектура проекта состоит из четырех базовых модулей: модуль формирования исходных данных, обработки данных, хранения данных и визуализация аналитики. Архитектура проекта представлена на рисунке 5.

Парсер получает данные через API HH и отправляет тело каждой вакансии в базу данных. Тело вакансии представлено в формате JSON. В качестве базы данных использовалась MongoDB. Далее csv файл загружается в HDFS. После чего с помощью Spark происходит обработка данных. Для визуализации аналитики была использована библиотека matplotlib.

HDFS — это распределенная файловая система, предназначенная для работы на стандартном оборудовании [4].

Некоторые особенности HDFS:

* Хранение больших объемов данных
* Большой размер блока
* Оптимизированная для параллельной потоковой работы с большими файлами
* Следование принципу: переместить вычисления проще, чем переместить данные [5].

Apache Spark — это многоязычный движок для разработки данных, науки о них и машинного обучения на одноузловых машинах или кластерах [6].

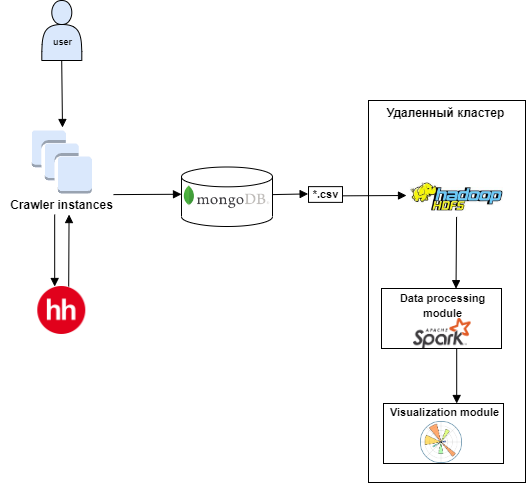


Рисунок 5 – Архитектура системы.

## Горизонтальное масштабирование

Система поддерживает горизонтальное масштабирование при сборе, обработке и хранении данных.

Парсер используемый в работе осуществлял параллельный сбор данных при помощи нескольких процессов. Каждый процесс производил сбор данных за определенный интервал времени. Чтобы избежать ограничения на количество запросов (10 запросов/сек/ip), сбор данных производился с нескольких proxy.

При обработке данных горизонтальное масштабирование осуществляется посредством увеличения числа Spark worker-ов.

HDFS обеспечивает горизонтальное масштабирование посредством распределения данных на множество серверов. Файл разбивается на блоки, которые хранятся на разных datanode с заданным коэффициентом репликации.

## Особенности созданного ПО

Базовые параметры GET запроса /vacancies, которые были использованы в работе: date\_from и date\_to. С помощью этих параметров можно выставить интервал, в котором будет происходить поиск вакансий. Данные параметры округляют указанное значение до ближайших пяти минут. Это означает, что минимальный шаг, за который можно получать результаты, будет равен пяти минутам.

Чтобы обойти ограничение на количество возвращаемых результатов в одном запросе (2000 вакансий), был использован переменных шаг. Стандартный шаг, за который производился сбор вакансий, равняется 30 мин. Если в одном из промежутков было найдено более 2000 вакансий, то данный промежуток дополнительно разбивался на более мелкие с шагом в 5 мин. Также парсер использует многопроцессорность, чтобы ускорить сбор данных.

## Инфраструктура разработки

В качестве языка программирования для реализации проекта был выбран Python версии 3.10.

Для управления изменениями в исходном коде проекта, а также для обеспечения эффективного сотрудничества разработчиков была использована система контроля версий Git.

Данные хранятся в нереляционной СУБД MongoDB в виде JSON-документов. MongoDB — документоориентированная система управления базами данных, не требующая описания схемы таблиц. Считается одним из классических примеров NoSQL-систем, использует JSON-подобные документы и схему базы данных. Написана на языке C++ [7].

В качестве интегрированной среды разработки был выбран PyCharm.

## Представление конечного вида дистрибутива

Дистрибутив представлен в виде веб приложения и состоит из следующих файлов:

* work.py - осуществляет сбор данных
* frame.py - осуществляет обработку данных
* requirements.txt
* Dockerfile
* app.py
* docker-compose.yaml

Готовый дистрибутив находится в открытом доступе на GitHub. Github: https://github.com/IlyaZawyalow/bigdata

# **Результаты**

Первоначально было поставлено 8 аналитических задач, которые в ходе работы были распределены на 5 групп.

## **Сравнительный анализ предлагаемых заработных плат в IT-сфере**

Задачи, относящиеся к этой группе, направлены на анализ заработных плат в IT-сфере в зависимости от различных критериев: города, профессиональной роли и типа занятости.

Анализ медианных заработных плат в зависимости от города (рисунок 6), как и ожидалось, показал, что наибольшие медианные заработные платы наблюдаются в Москве (121800 руб.) и Санкт-Петербурге (90000 руб.). В топ также попали Новосибирск (75000 руб.), Екатеринбург (70452.5 руб.), Владивосток (70000 руб.) и другие.

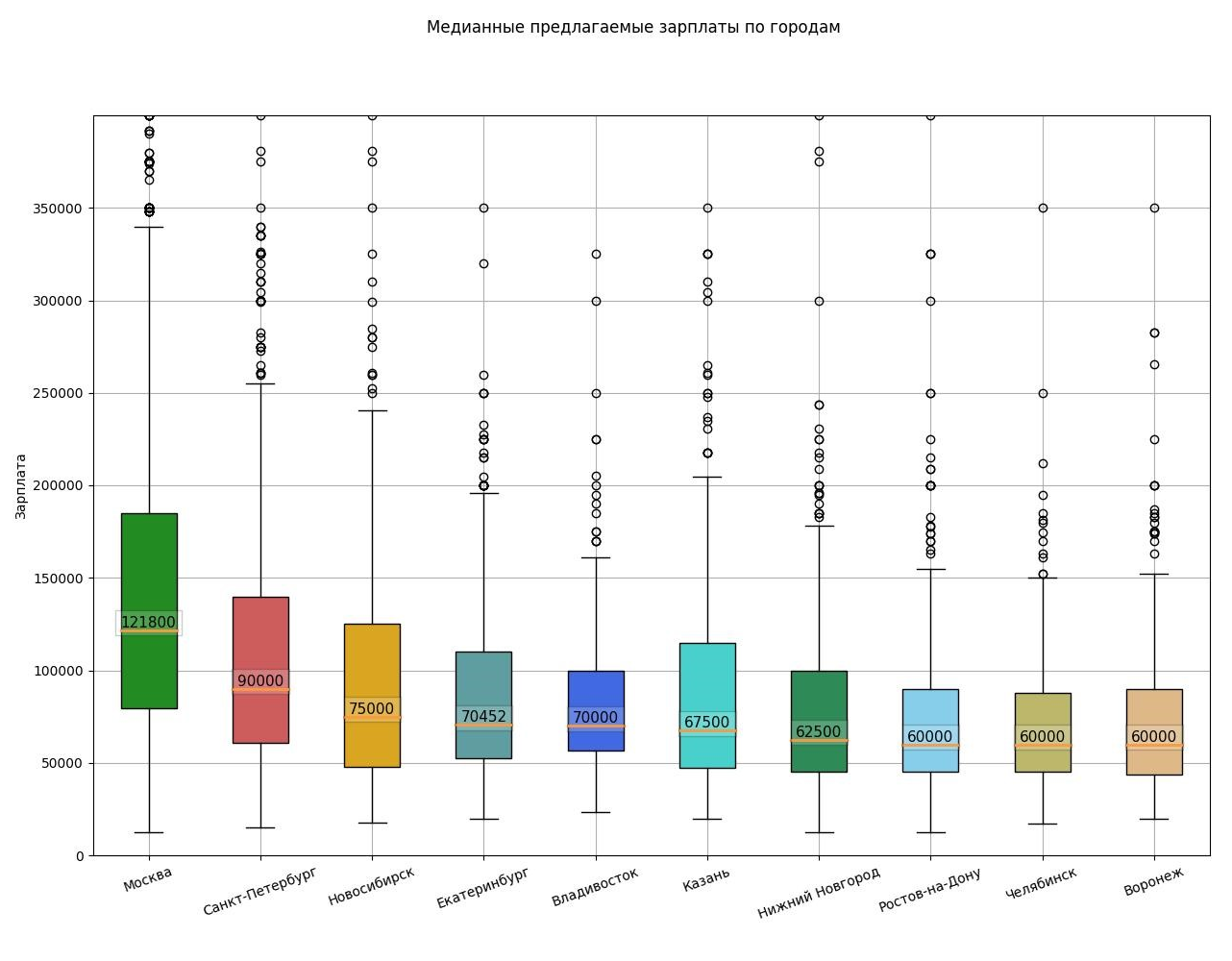


Рисунок 6 – Медианные предлагаемые зарплаты по городам

На рисунке 7 представлены медианные заработные платы в зависимости от профессиональной роли. Из графика видно, что наименьшие медианные заработные платы имеют дизайнер (57500 руб.) и специалист по информационной безопасности (70000 руб.), в то время как наибольшие медианные зарплаты имеют DevOps-инженер (200 тысяч руб.) и Team Lead (172 тысячи руб.).

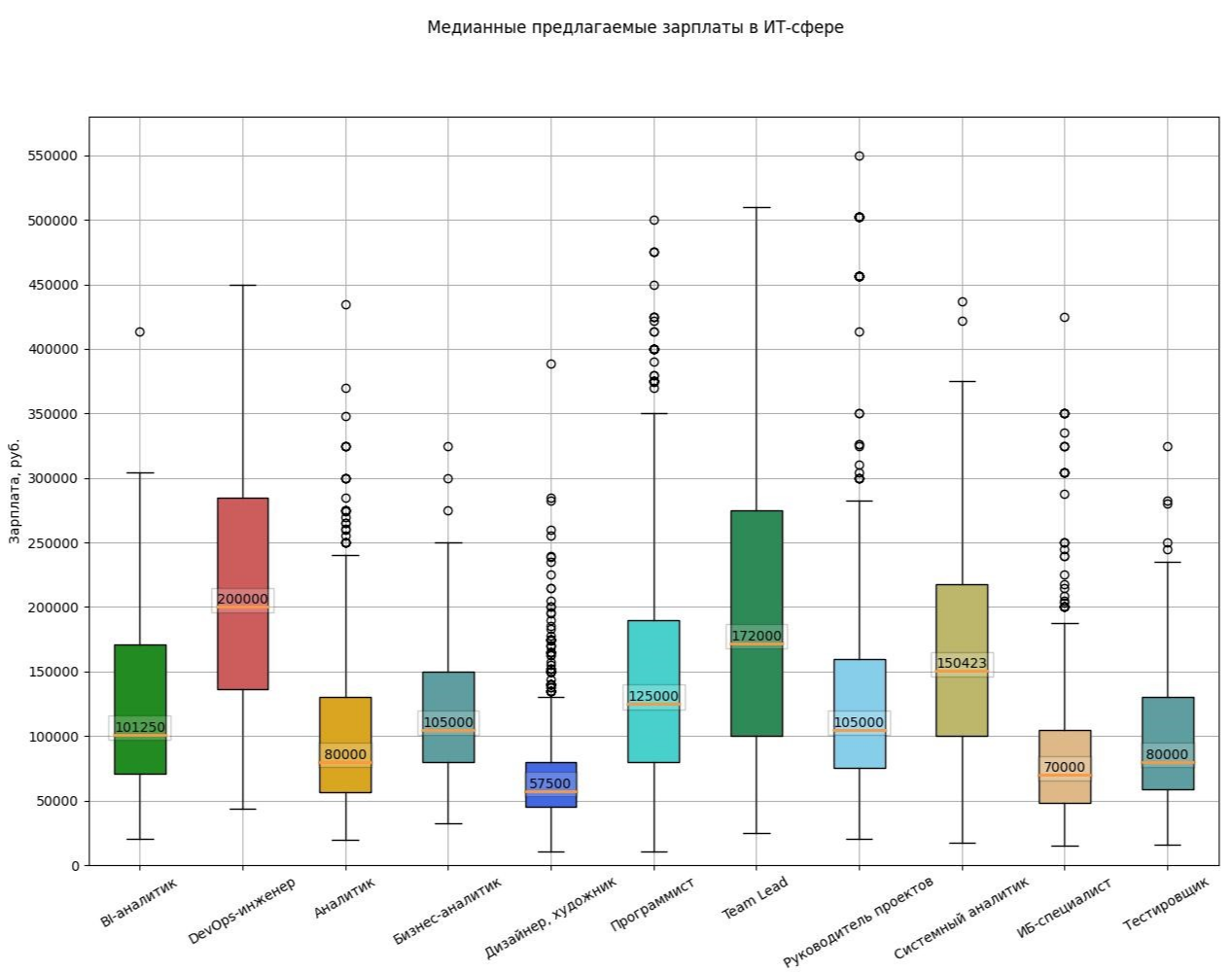


Рисунок 7 – Медианные предлагаемые зарплаты в ИТ-сфере

Также был проведен сравнительный анализ медианных заработных плат у вакансий с разным типом занятости (рисунок 8). Из графика видно, что частичная занятость имеет наименьшую медианную заработную плату (31500 руб.), причем она более чем в два раза ниже, чем зарплата при полной занятости (72500 руб.). Это может говорить о том, что спрос на работников с частичной занятостью ниже, что влияет на уровень заработной платы.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание

Рисунок 8 – Медианные предлагаемые зарплаты в ИТ-сфере в зависимости от типа занятости

## **Исследование распределения требуемого опыта от соискателей и анализ оплаты труда в зависимости от опыта.**

Для понимания ситуации на рынке труда в ИТ-сфере полезно также узнать, как распределён рынок ИТ-сферы по требуемому опыту.

На рисунке 9 изображено распределение требуемого опыта работы в ИТ-сфере. Из графика видно, что наименьшее количество вакансий требует от соискателей опыта работы более шести лет.

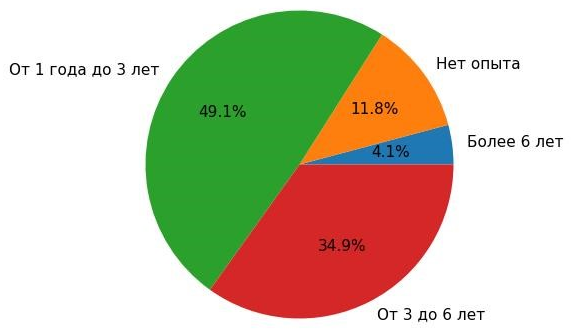
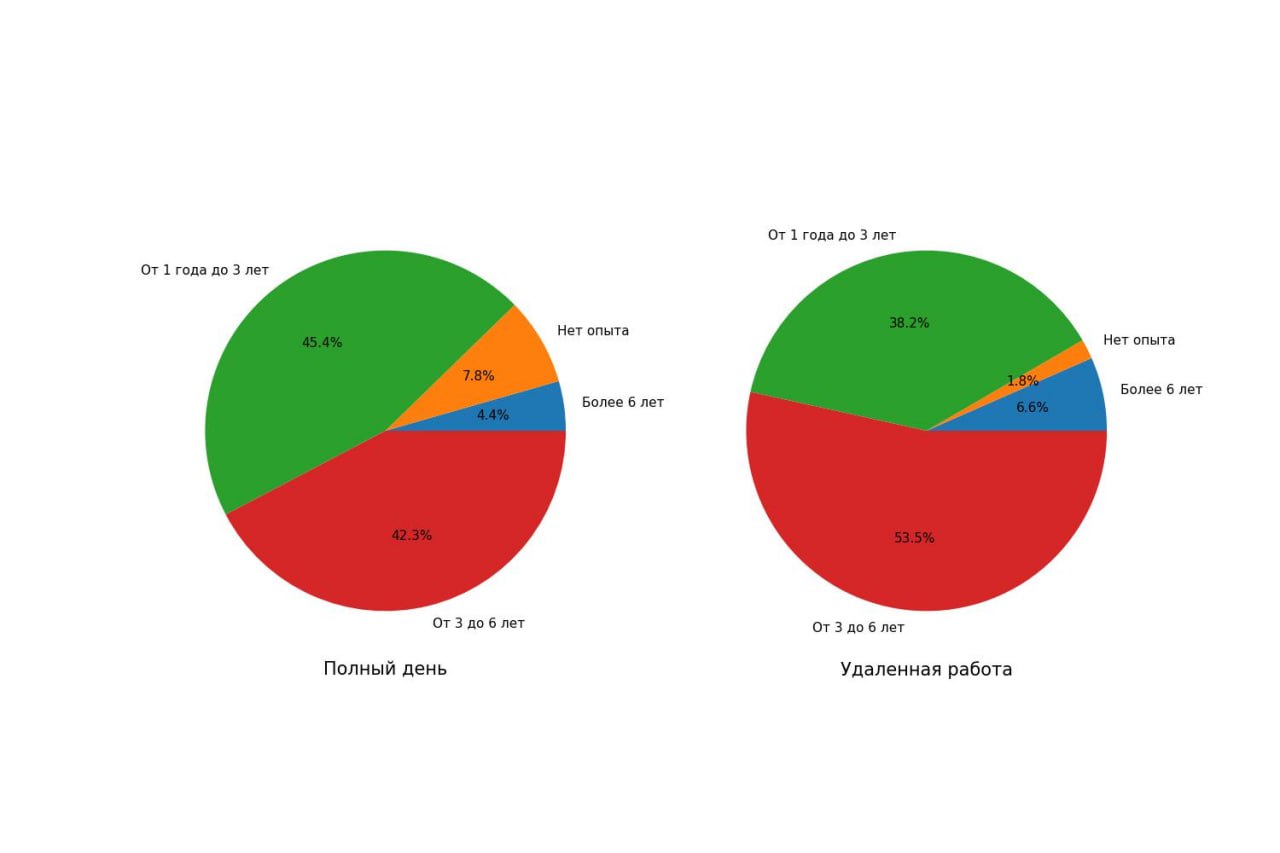


Рисунок 9 – Распределение опыта

На рисунке 10 отображено распределение требуемого опыта в зависимости от графика работы для профессии "программист, разработчик". Из графика видно, что для удаленной работы обычно требуются более опытные сотрудники: более половины таких вакансий предполагают наличие опыта работы от трех лет.

Рисунок 10 – Распределение требуемого опыта в зависимости от графика работы для профессии "программист, разработчик"

Зарплаты достаточно сильно зависят от навыков и опыта работы кандидатов, и ИТ-сфера в данном случае не исключение. На рисунке 11 изображена зависимость заработной платы от опыта работы. Как видно из графика, медианная заработная плата для начинающих в ИТ составляет до 47500 руб., а у специалистов с опытом работы более шести лет доходит до 165 тысяч рублей. Кроме того, медианные заработные платы у специалистов с опытом работы от одного до трёх лет и специалистов с опытом от 3 до 6 лет отличаются более чем в два раза.

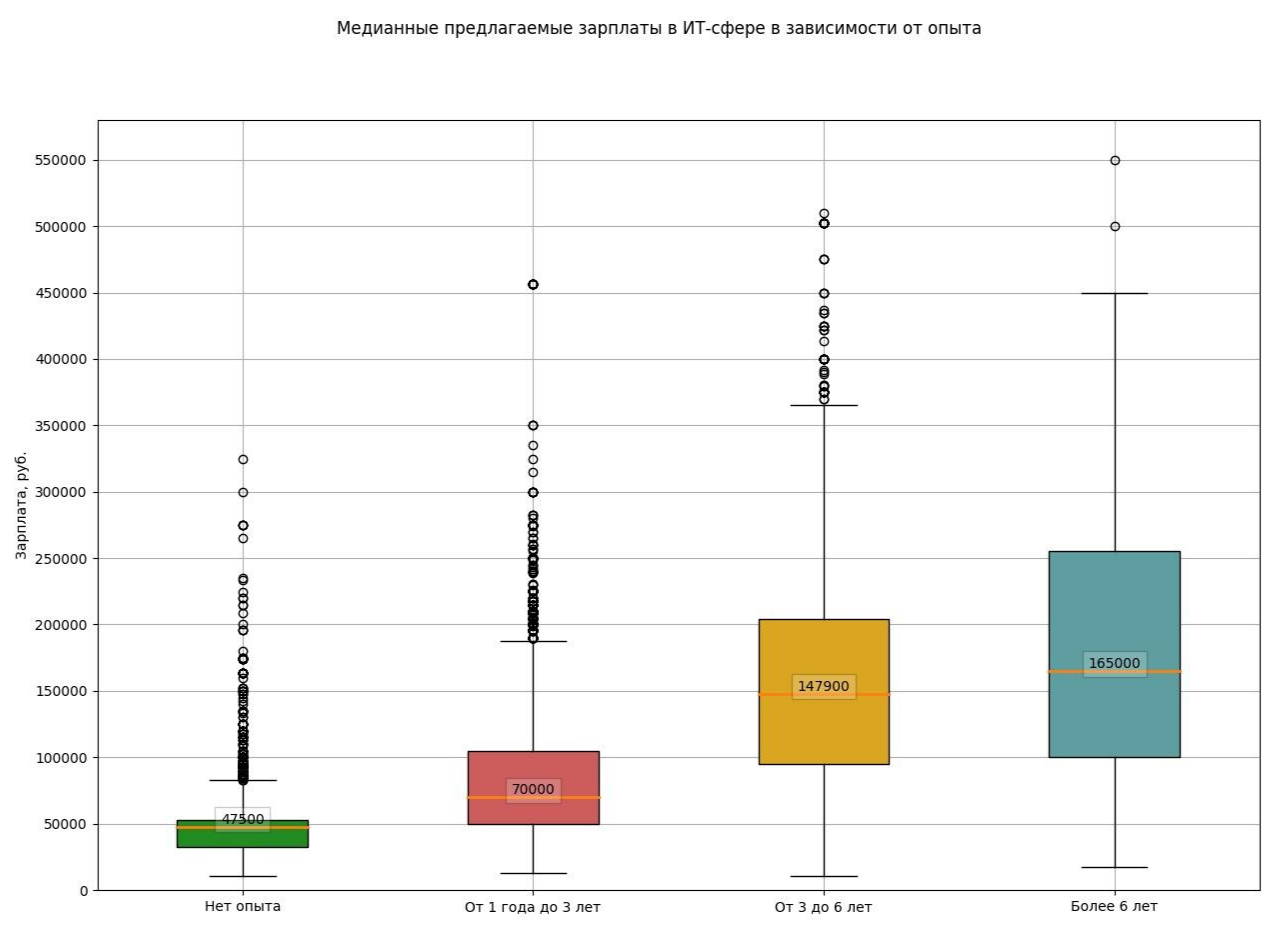


Рисунок 11 – Медианные предлагаемые зарплаты в ИТ-сфере в зависимости от опыта

## **Определение топ-работодателей**

Задачи, относящиеся к этой группе, направлены на определение компаний, наиболее активно размещающих вакансии в IT-сфере, и анализ их потребностей в персонале, ключевых навыках и опыте работы.

Компании, наиболее активно размещающие вакансии, представлены в таблице 2.

Таблица 2 – Компании, наиболее активно размещающие вакансии

|  |  |
| --- | --- |
| Работодатель | Количество вакансий |
| СБЕР | 1341 |
| Ozon | 889 |
| МТС | 606 |
| Тинькофф | 531 |
| Иннотех, Группа компаний | 512 |
| Яндекс | 367 |
| ИК СИБИНТЕК | 300 |
| Газпром нефть | 295 |
| Банк ВТБ (ПАО) | 248 |
| Первый Бит | 234 |

На рисунке 12 представлено распределение опыта работы в крупных компаниях. Заметно, что вне зависимости от компании наибольшее число вакансий требует опыт работы от 1 до 3 лет и от 3 до 6 лет.

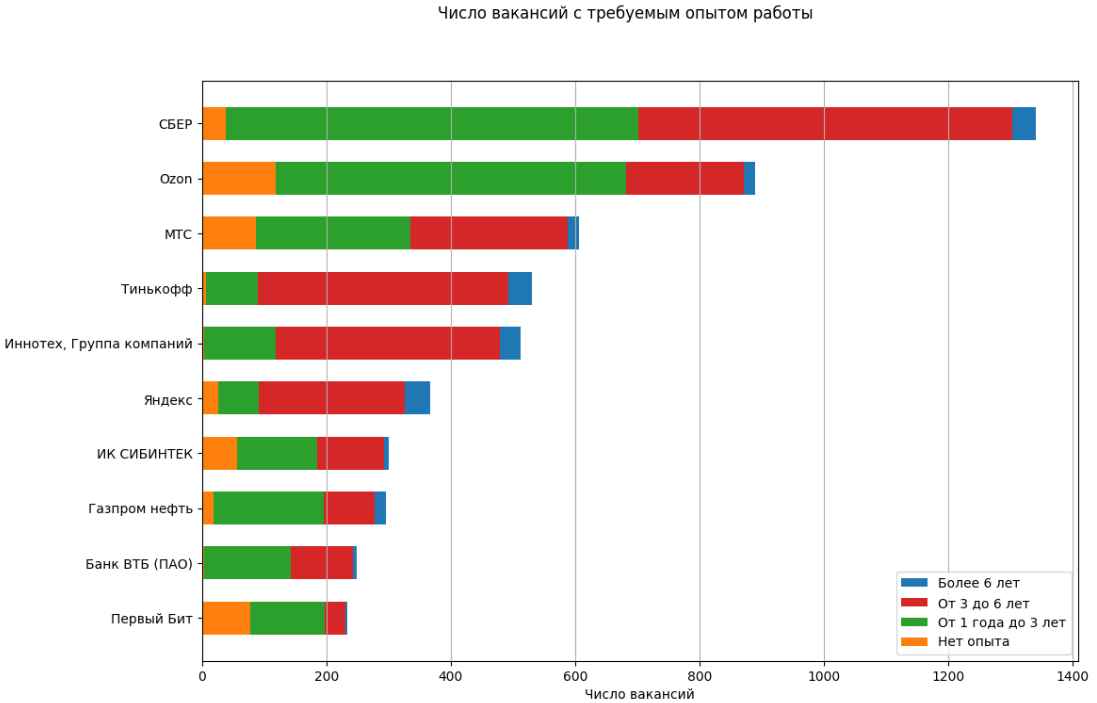


Рисунок 12 – Число вакансий с требуемым опытом работы

На рисунке 13 представлено распределение языков программирования в крупных компаниях. В качестве языков для анализа были выбраны Python, Java, JavaScript и Golang. Из графика видно, что в большинстве вакансий в качестве требуемых языков указаны Python и Java.

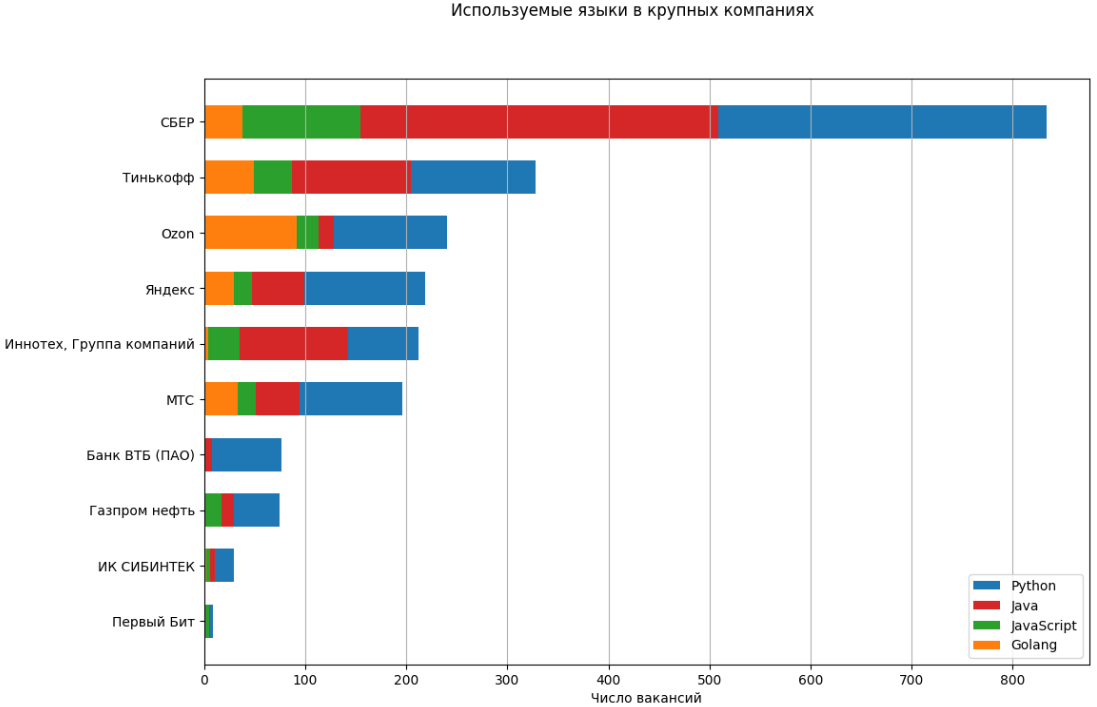
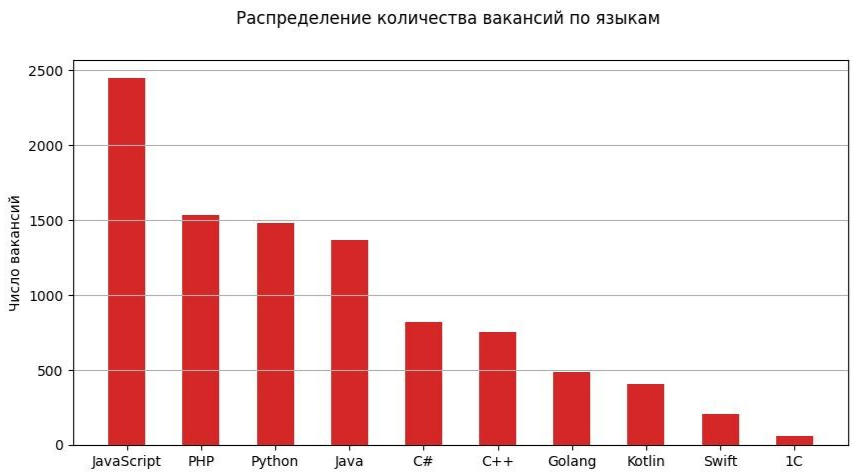


Рисунок 13 – Используемые языки в крупных компаниях

## **Определение наиболее востребованных навыков для профессии программиста разработчика**

Решение данной задачи представлено на рисунке 14. Из графика видно, что наиболее часто работодатели указывают в требованиях JavaScript, PHP и Python.

Рисунок 14 – Распределение количества вакансий по языкам

## **Тепловая карта IT-вакансий**

Так как в теле вакансии указаны координаты места работы, получилось перенести данные на географическую карту (рисунок 15). Красными точками указаны станции метро.



Рисунок 15 – Тепловая карта IT-вакансий в Санкт-Петербурге

## **ЗАКЛЮЧЕНИЕ**

В процессе выполнения данной работы были выполнены следующие этапы:

* Выполнено описание процесса сбора данных
* Описаны нюансы работы с API HH
* Выполнено описание программного решения
* Приведены результаты анализа полученных данных

По результатам, полученным в работе, можно сделать вывод, анализ данных используя API HH позволил выявить тренды и предпочтения работодателей, что может быть полезно для соискателей.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Официальный сайт HeadHunter [Электронный ресурс] URL: https://kirishi.hh.ru/article/28/ (дата обращения: 15.11.2023).
2. Официальный сайт HeadHunter API [Электронный ресурс] URL: https://api.hh.ru/ (дата обращения: 15.11.2023).
3. Официальная документация сервиса HeadHunter API. [Электронный ресурс] URL: https://api.hh.ru/openapi/redoc#tag/ (дата обращения: 15.11.2023).
4. Официальный сайт Apache Hadoop [Электронный ресурс] URL: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html/ (дата обращения: 15.11.2023).
5. Лекционные материалы по дисциплине «Наука о данных и аналитика больших объемов данных» – И. В. Никифоров. Институт компьютерных наук и технологи. Высшая школа программной инженерии. СПБПУ. – 2023.
6. Официальный сайт Apache Spark [Электронный ресурс] URL: https://spark.apache.org/ (дата обращения: 15.11.2023).
7. MongoDB — Википедия [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/MongoDB (дата обращения: 15.11.2023).