Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет информатики и радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра информатики

Отчет по лабораторной работе
по курсу «Технологическая платформа по управлению большими данными»
на тему «**Обработка данных о вакансиях**»

Выполнил студент группы 956241: Зязюлькин С.П.

Проверил: Стержанов М.В.

Минск, 2020

Постановка задачи

Задача заключается в анализе информации о вакансиях в России. В частности, делается акцент на объёме анализируемых данных (должно быть минимум 100000 вакансий для анализа) и на сравнении данных по России с данными по Беларуси. В рамках анализа также предполагается сравнить медианный и средний уровни предлагаемой в вакансиях заработной платы с официальной статистикой.

План выполнения работы

1. Найти источник данных о вакансиях в России, содержащий не менее 100000 вакансий.
2. Реализовать выгрузку данных из источника.
3. Выполнить обработку загруженных данных.
4. Провести анализ обработанных данных и выполнить сравнение с данными по Беларуси.
5. Сделать выводы.

Получение данных

В качестве источника данных о вакансиях в Беларуси выбран сайт https://russia.trud.com. Сайт содержит более миллиона вакансий, однако для каждого результата поиска выдаётся не более 3000 вакансий. Поэтому просто выгрузить с главной страницы более 3000 вакансий не представляется возможным.

Для каждой вакансии выгружаются следующие данные:

- Предлагаемая должность.
- Диапазон заработной платы.
- Место.
- Наниматель.
- Краткое описание вакансии.

Стоит отметить, что для некоторых вакансий может отсутствовать часть данных, т.е. не для каждой вакансии все эти данные заполнены.

Выгрузка данных осуществлялась с использованием языка программирования Python и фреймворка Scrapy.

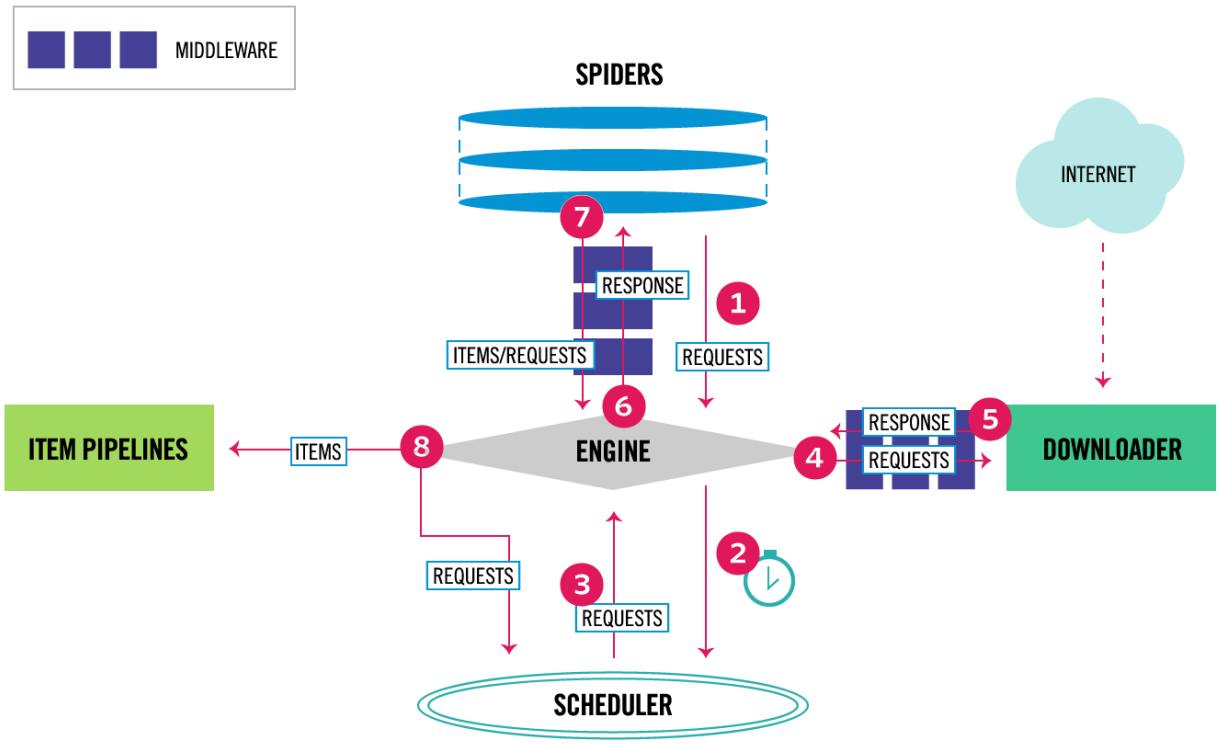

Рисунок 1 – Архитектура фреймворка Scrapy

Для выгрузки данных необходимо было реализовать паука (Spider), который отвечает за парсинг выгружаемых html-страниц, а также указывает, какие страницы (URL) необходимо парсить.

В этот раз простой парсинг вакансий с главной страницы не представляется возможным из-за ограничения в 75 страниц по 40 вакансий, т.е. ограничения в 3000 вакансий. Поэтому было принято решение действовать следующим образом. На главной странице находятся ссылки на вакансии по конкретным городам, что позволит выгрузить до 3000 вакансий для каждого города, что, с учётом числа городов, позволит выгрузить более 200000 вакансий.

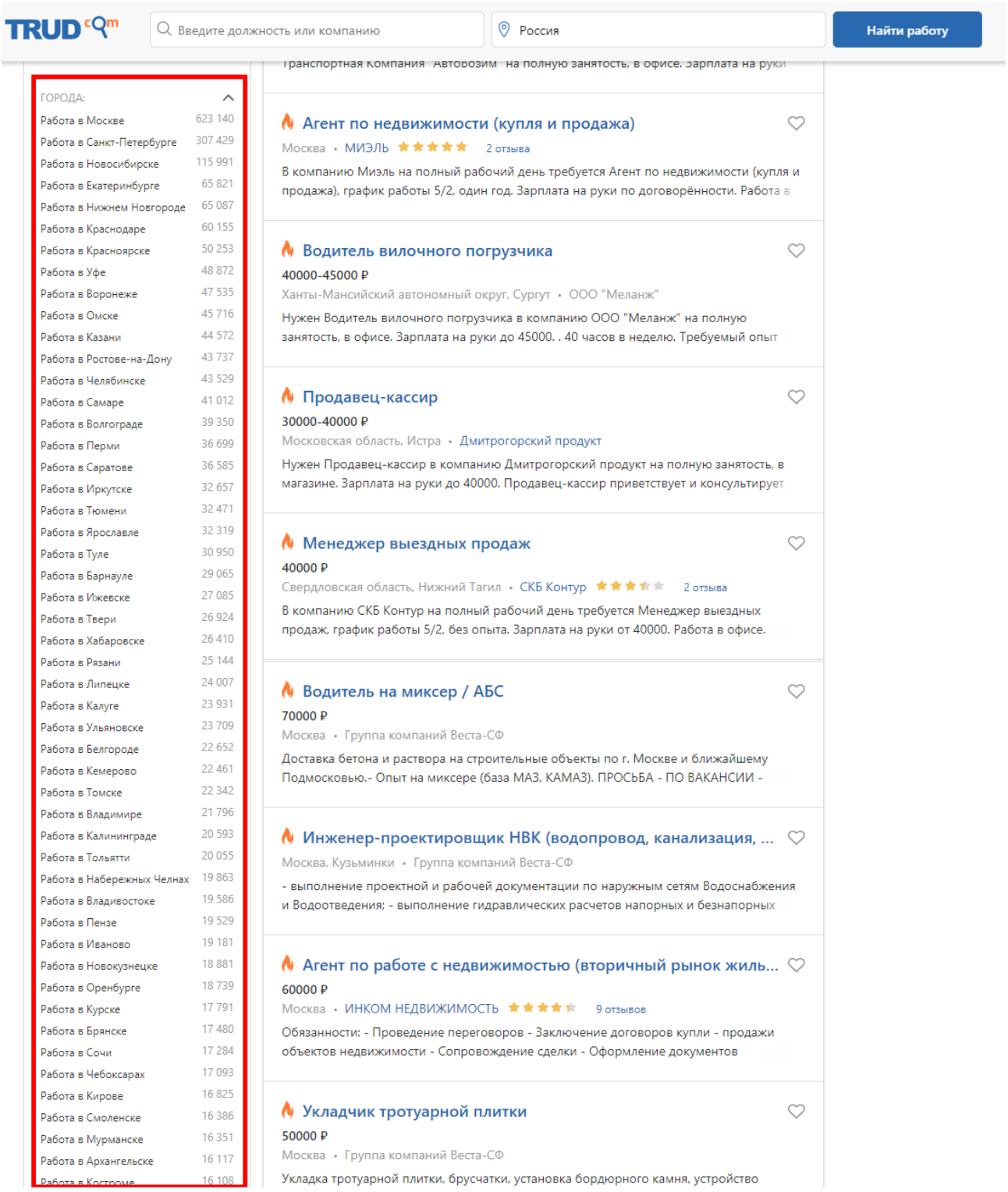

Рисунок 2 – Ссылки на вакансии по городам

Другая трудность, с которой быстро сталкиваешься в процессе выгрузки данных, состоит в том, что сайт быстро распознаёт в пауке робота и начинает возвращать ему капчу, чтобы подтвердить, что обращения делает не робот.

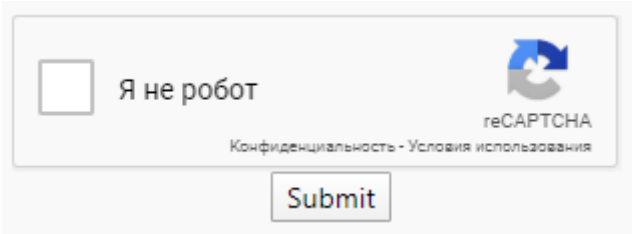

Рисунок 3 – Защита от роботов

Чтобы обойти эту защиту, была выполнена настройка паука, чтобы он делал меньше запросов и интервал между запросами был больше. В частности, паук слал не больше одного запроса за раз, перед очередным запросом он делал паузу в секунду. В результате, сайт крайне редко включал проверку на робота, а в тех редких случаях, когда это всё же происходило, проверка на робота решалась вручную, после чего паук спокойно продолжал выгрузку данных.

Код паука:

```
class JobsSpider(scrapy.Spider):
    name = 'jobs_spider'
    base_url = 'https://russia.trud.com'
    start_urls = [base_url]

    def parse(self, response):
        print(response)

        if response.url == JobsSpider.base_url:
            for region in response.css('div[class="sidebar-item
region-4-unit"] li'):
                region_url = region.css('a').attrib['href']
                for page in reversed(range(1, 76)):
                    yield scrapy.Request(
                        region_url if page == 1 else
f'{region_url}page/{page}',
                        callback=self.parse
                    )
        else:
            for job in response.css('div[class*="card"]'):
                yield JobsSpider._to_json(job)

    @staticmethod
    def _to_json(job):
        name = job.css('a[class="item-
link"]::text').extract_first()
        salary =
job.css('span[class*="salary"]::text').extract_first()
        place = job.css('span[class*="geo-
location"]::text').extract_first()

        employer = job.css('span[class*="institution"]
span::text').extract_first()
        if not employer:
```

```
            employer = job.css('a[class="company-link"]
span::text').extract_first()

        description = job.css('div[class="item-
description"]::text').extract_first()
        if not description:
            description = job.css('div[class="item-
info"]::text').extract_first()

        if name:
            return {
                'name': name,
                'salary': salary,
                'place': place,
                'employer': employer,
                'description': description,
            }

jobs_file = 'jobs.json'

if os.path.isfile(jobs_file):
    os.remove(jobs_file)

process = CrawlerProcess({
    'LOG_LEVEL': logging.INFO,
    'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT
5.1)',
    'FEED_FORMAT': 'json',
    'FEED_URI': jobs_file,
    'FEED_EXPORT_ENCODING': 'utf-8',
    'CONCURRENT_ITEMS': 1,
    'CONCURRENT_REQUESTS': 1,
    'DOWNLOAD_DELAY': 1
})

process.crawl(JobsSpider)
process.start()
```

Выгрузка заняла несколько часов. Было выгружено 204800 вакансий, что почти в 35 раз больше, чем в рамках анализа вакансий по Беларуси.

Выгруженные данные были сохранены в файл в формате json. Итоговый размер файла составил 110 МБ, что почти в 67 раз больше, чем в рамках анализа вакансий по Беларуси.

Пример данных одной вакансии в формате json: {"name": "WEB-программист", "salary": "200000 ₽", "place": "Москва", "employer": "Работа вакансии рф", "description": " Создание и поддержка веб-сервиса на Drupal 7, Laravel. Разработка модулей Drupal. Разработка серверной и клиентской части сайта. Разработка личного кабинета для клиентов. Написание и оптимизация сложных запросов к ... "}.

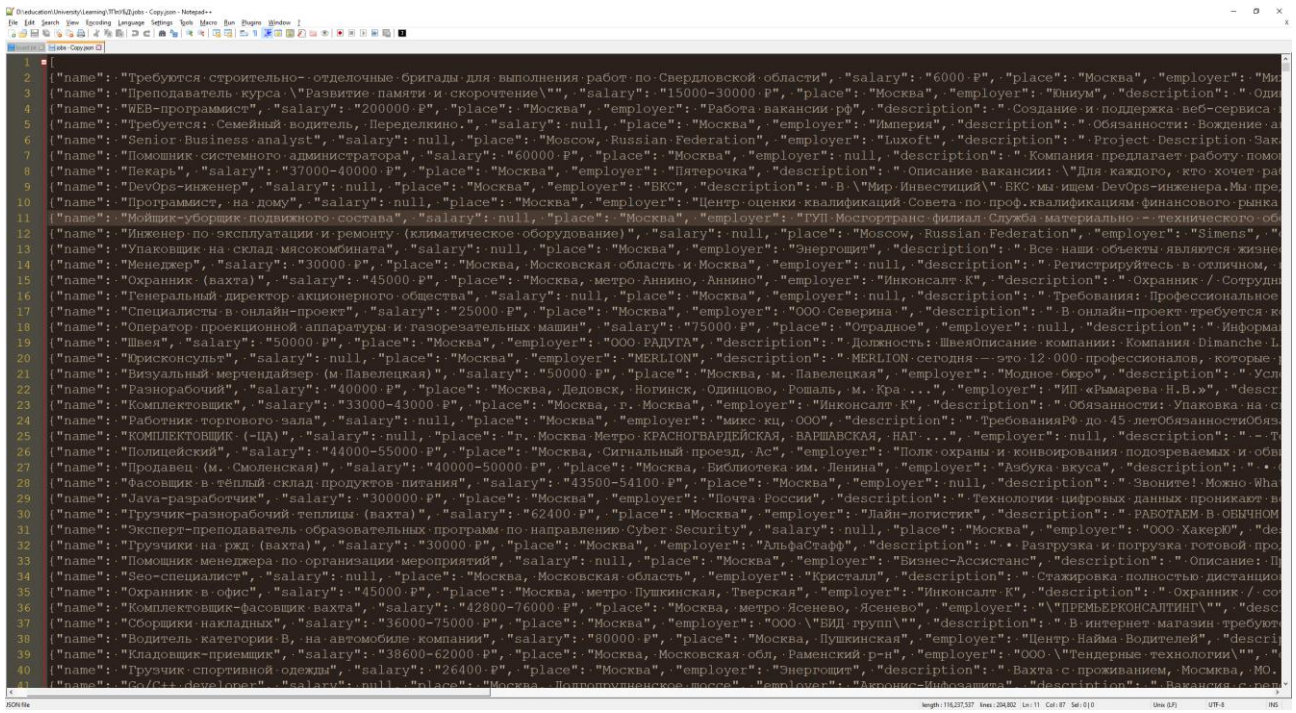

Рисунок 4 – Файл с данными

Обработка выгруженных данных

Обработка выгруженных данных выполнялась на языке Python с использованием стандартной библиотеки.

Действия, выполняемые во время обработки:

1. Если в вакансии отсутствует предлагаемая должность, то она заменяется на "Не указано". Также обрезаются пробельные символы в поле.
2. Из поля "заработная плата" при помощи регулярных выражений удаляются лишние пробельные символы и лишний текст, а затем извлекается верхняя и нижняя граница заработной платы (пример формата до извлечения границ: 3000-5000 ₽). Одна или обе границы могут быть не указаны. Также проверяется, что валюта указана и совпадает с российским рублём.
3. Если в вакансии отсутствует место (адрес), то оно заменяется на "Не указано". Также обрезаются пробельные символы в поле.
4. Если в вакансии отсутствует наниматель, то он заменяется на "Не указан". Также обрезаются пробельные символы в поле.
5. Если в вакансии отсутствует краткое описание, то оно заменяется на "Не указано". Также обрезаются пробельные символы в поле.
6. Для всех полей, содержащих категориальные данные, формируется множество допустимых значений. К таким полям относятся следующие: предлагаемая должность, наниматель.

Категориальные данные

В данном разделе для выгруженных данных приводится список допустимых значений категориальных полей.

- Предлагаемая должность (10 первых значений, общее число: 53685): 'Грузчик - подсобник на склад', 'Мойщик автомобилей в дилерском центре', 'Продавец-кассир_Подработка (Курск, ул Песковская 3-я, 1)', 'Оператор связи (г. Тверь)', 'Механик перегрузочных машин (по погрузочно-разгрузочным механизмам)', 'Сиделка в пансионат для пожилых', 'Технический специалист на склад', 'Инженер по подключению корпоративных клиентов', 'Грузчик (ул. 9 мая 62)', 'Специалист дистанционного клиентского обслуживания и продаж'.
- Наниматель (10 первых значений, общее число: 33331): 'Министерство инвестиционного развития Забайкальского края', 'Такси "Мак - Авто"', 'Ольга-С', 'ООО ЛуидорГарантия-КАЗАНЬ (АВТОЦЕНТР ГАЗ-ЛУИДОР)', '(ООО "ППФ Страхование жизни") Индивидуальный предприниматель Абдулина Алия Рашидовна', 'ГУВ МО "Московская областная ветеринарно-санитарная станция"', 'СПЕЦАВТОПРОМ', 'Земские Просторы', 'ООО "Комплекс "Дворец Молодежи"', 'Панацея'

- **Анализ данных**

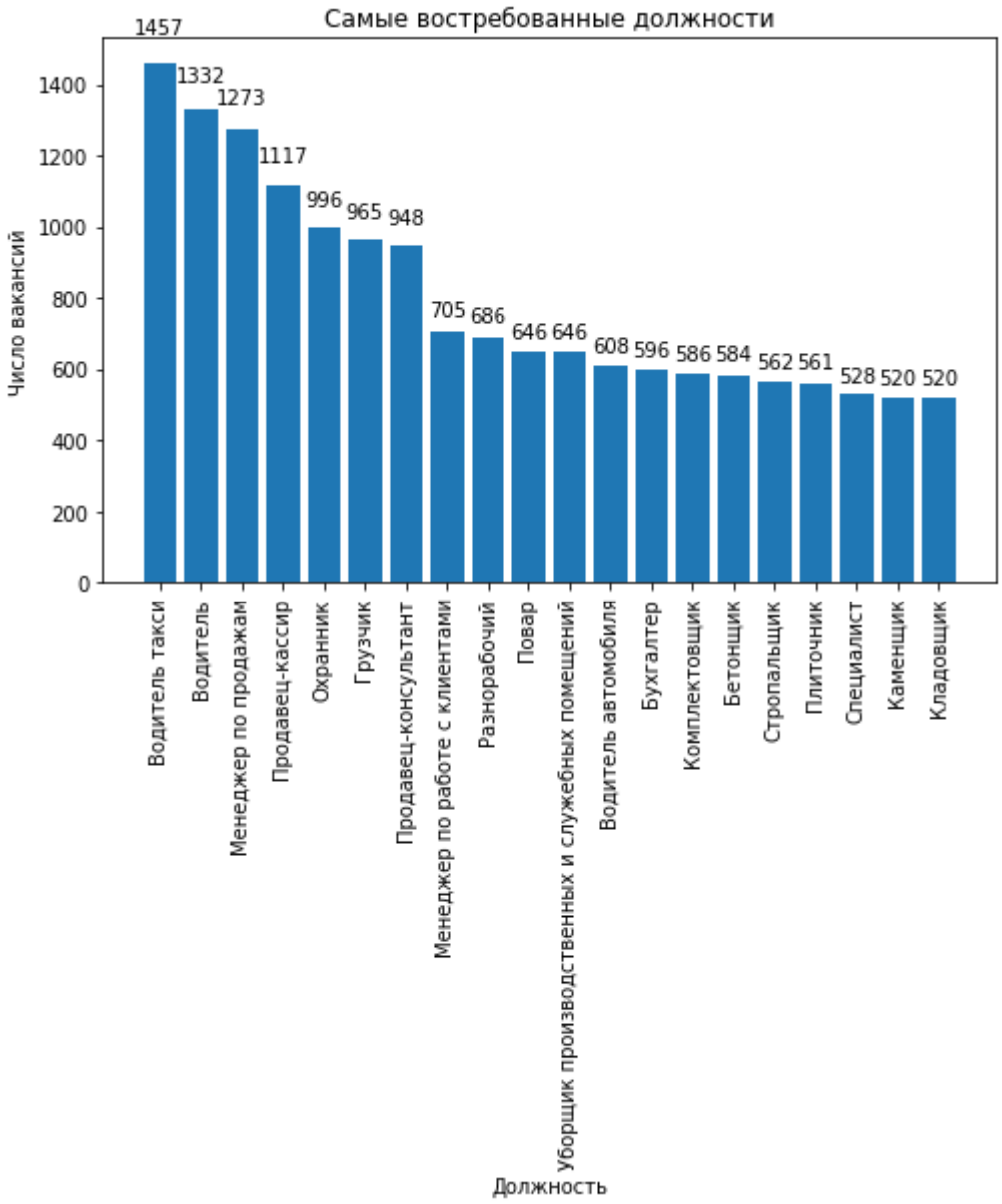

Рисунок 5 – Распределение вакансий по предлагаемым должностям (топ 20 должностей)

Большая часть вакансий представлена следующими должностями: водитель, менеджер, продавец, охранник, грузчик, разнорабочий, повар, уборщик, бухгалтер, комплектовщик, строитель.

Если сравнивать с данными по Беларуси, то там чаще всего требовались продавцы, повары и водители, что соответствует данным по России.

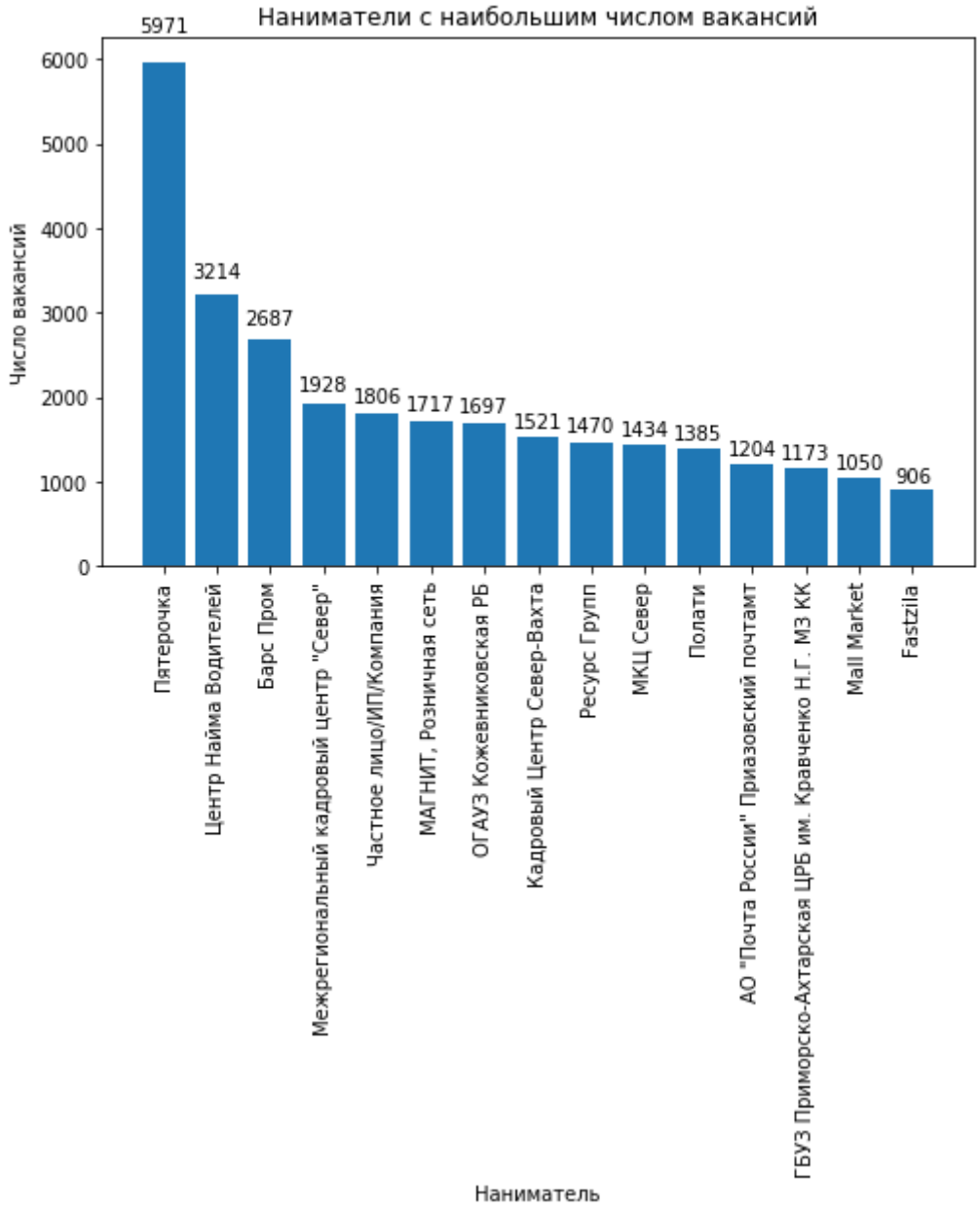

Рисунок 6 – Распределение вакансий по нанимателям (топ 20 значений)

Значительно превосходит других нанимателей по числу вакансий «Пятерочка». В топе нанимателей фигурируют сети магазинов и различного рода центры. Также стоит отметить большое число вакансий от частных лиц, ИП и компаний. В целом, данные соответствуют востребованным должностям.

Стоит отметить, что данные значительно отличаются от данных по Беларуси, где в топе нанимателей фигурировали застройщики.

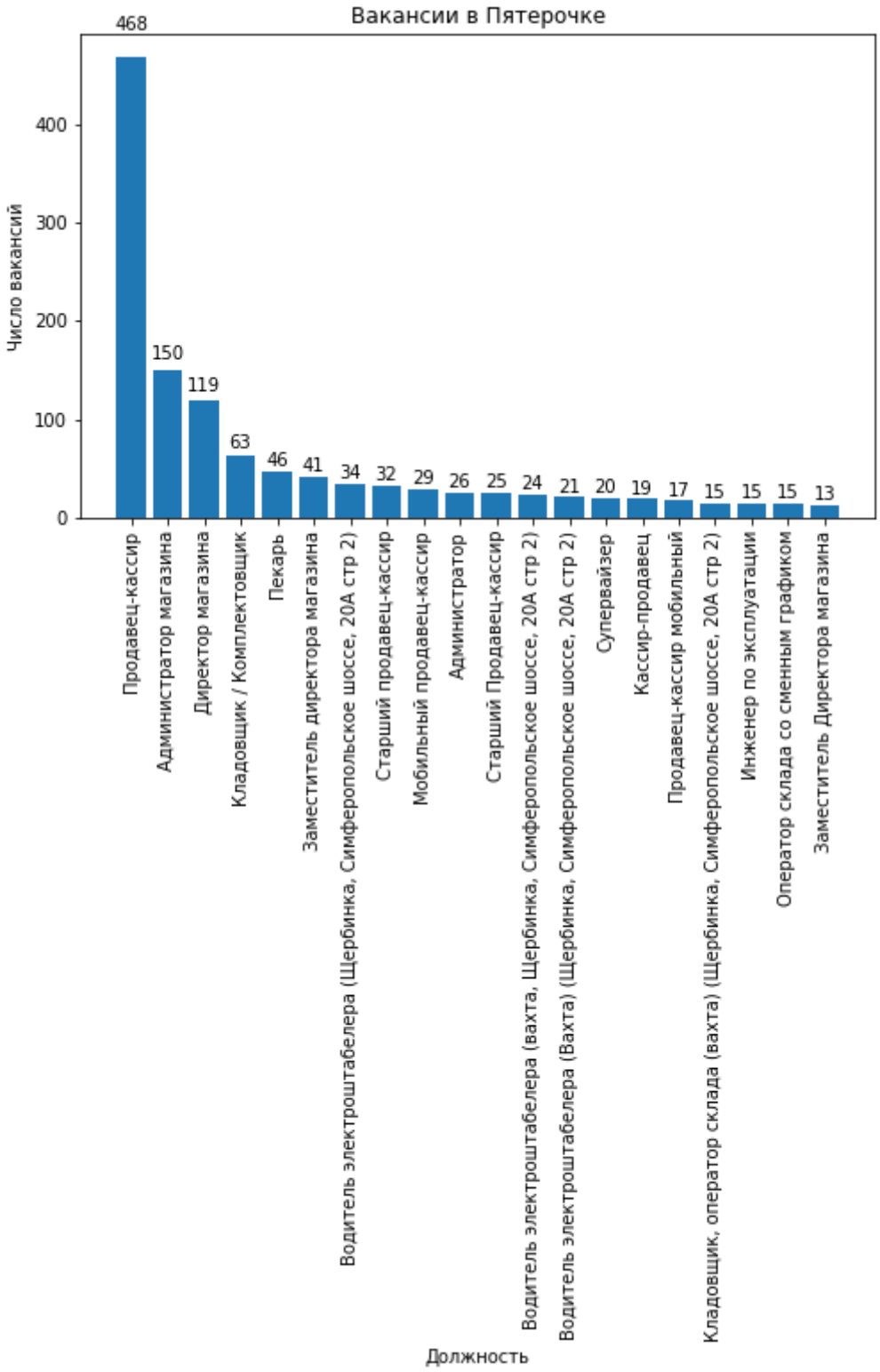

Рисунок 7 – Распределение вакансий по предлагаемым должностям в «Пятерочке» (топ 20 должностей)

Как и ожидалось, больше всего требуется продавцов. Удивляет, что требуется такое большое число директоров магазина и их заместителей.

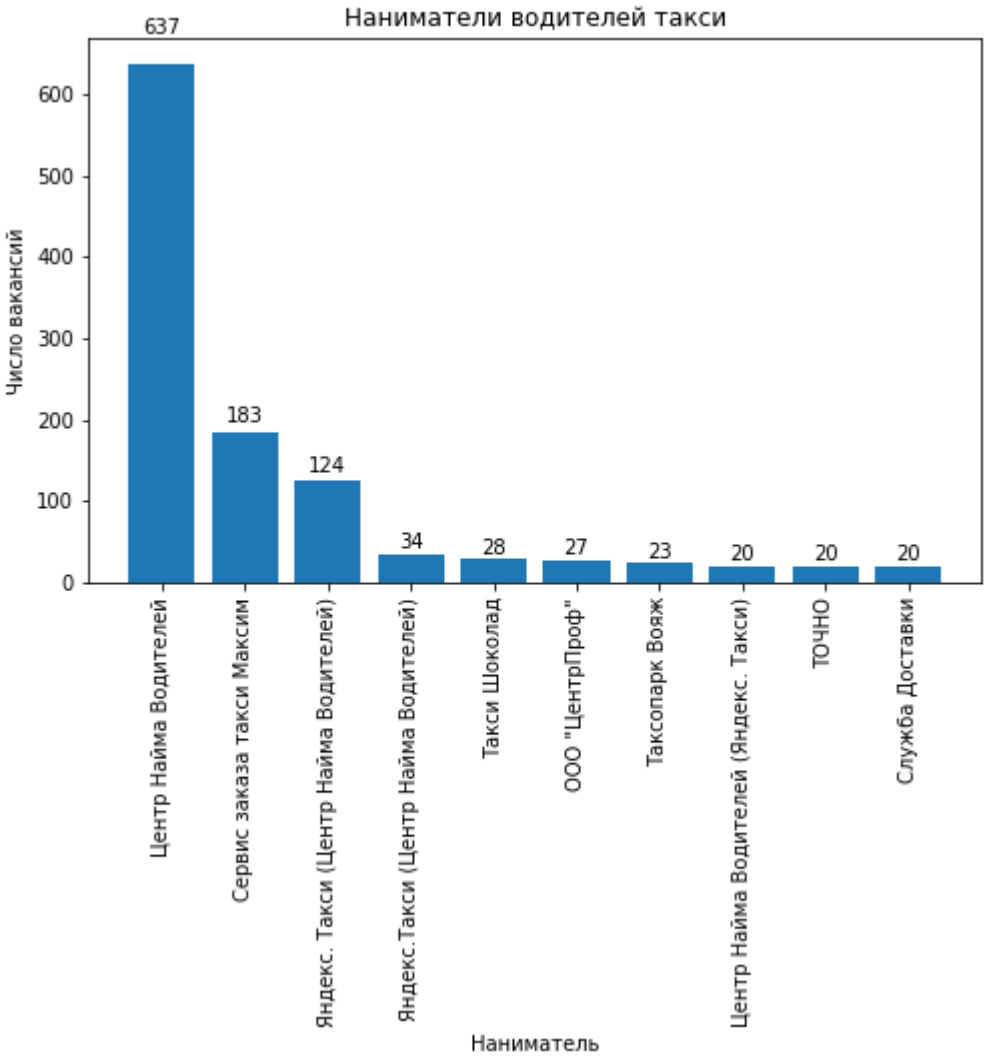

Рисунок 8 – Распределение вакансий «водитель такси» по нанимателям (топ 10 значений)

Совсем не удивляет, что подавляющее число водителей ищет «Центр Найма Водителей». Также стоит выделить среди нанимателей «Яндекс.Такси».

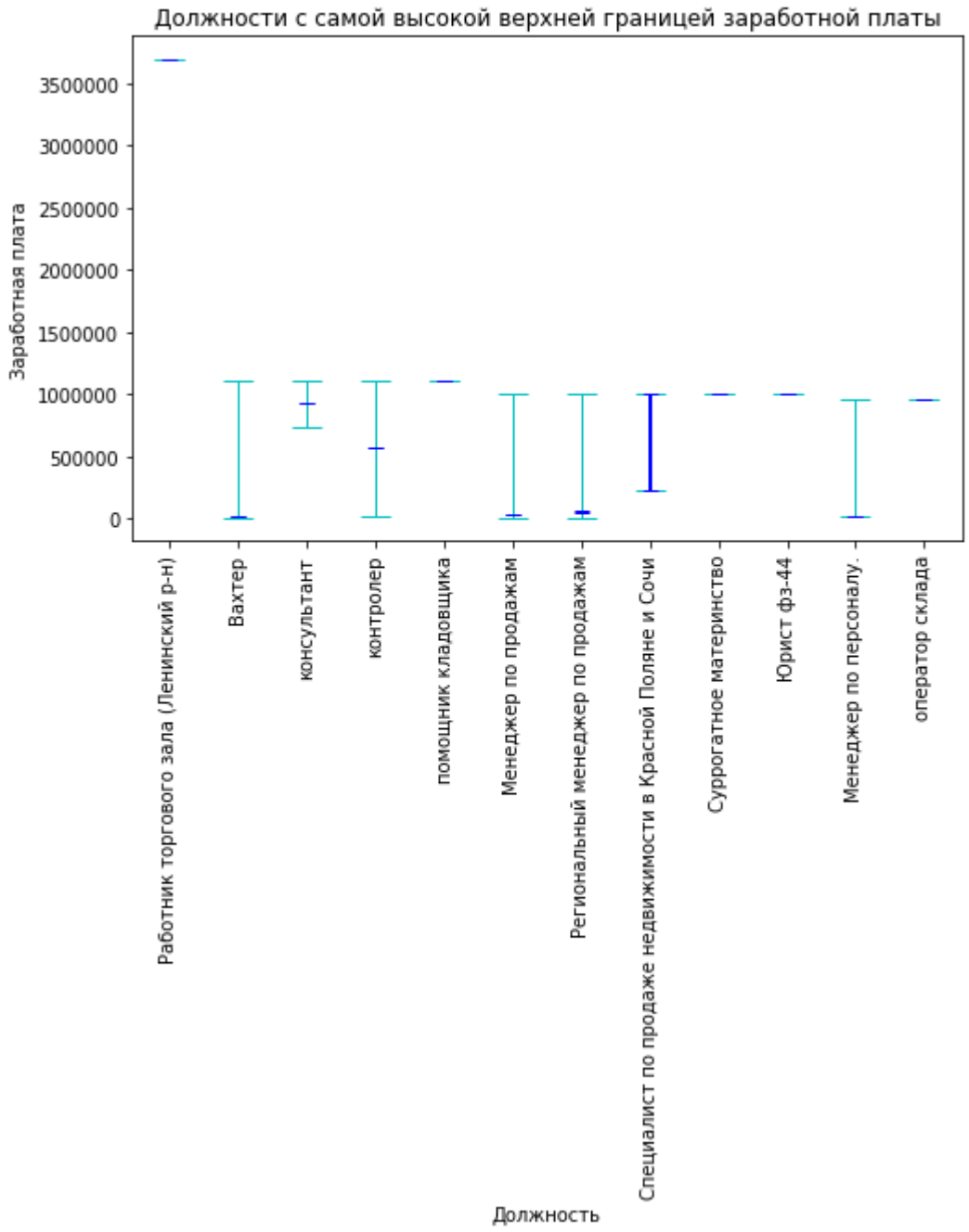

Рисунок 9 – Зависимость абсолютной верхней границы предлагаемой заработной платы от должности (топ 12)

Интересно видеть, что должности с самой высокой возможной заработной платой также имеют очень низкую нижнюю границу. Заработная плата вакансии "Вахтер" может варьироваться от 9500 до 1109880 ₽, т.е. верхняя граница отличается от нижней более чем в 100 раз. Похожая ситуация наблюдается и у многих других вакансий. Отметим, что абсолютно такая же ситуация наблюдалась и в данных по Беларуси. Предполагаю, что причина кроется в том, что наниматели ставят очень большую верхнюю границу заработной платы, чтобы просто привлечь внимание. В реальности такую заработную плату работник никак не получит.

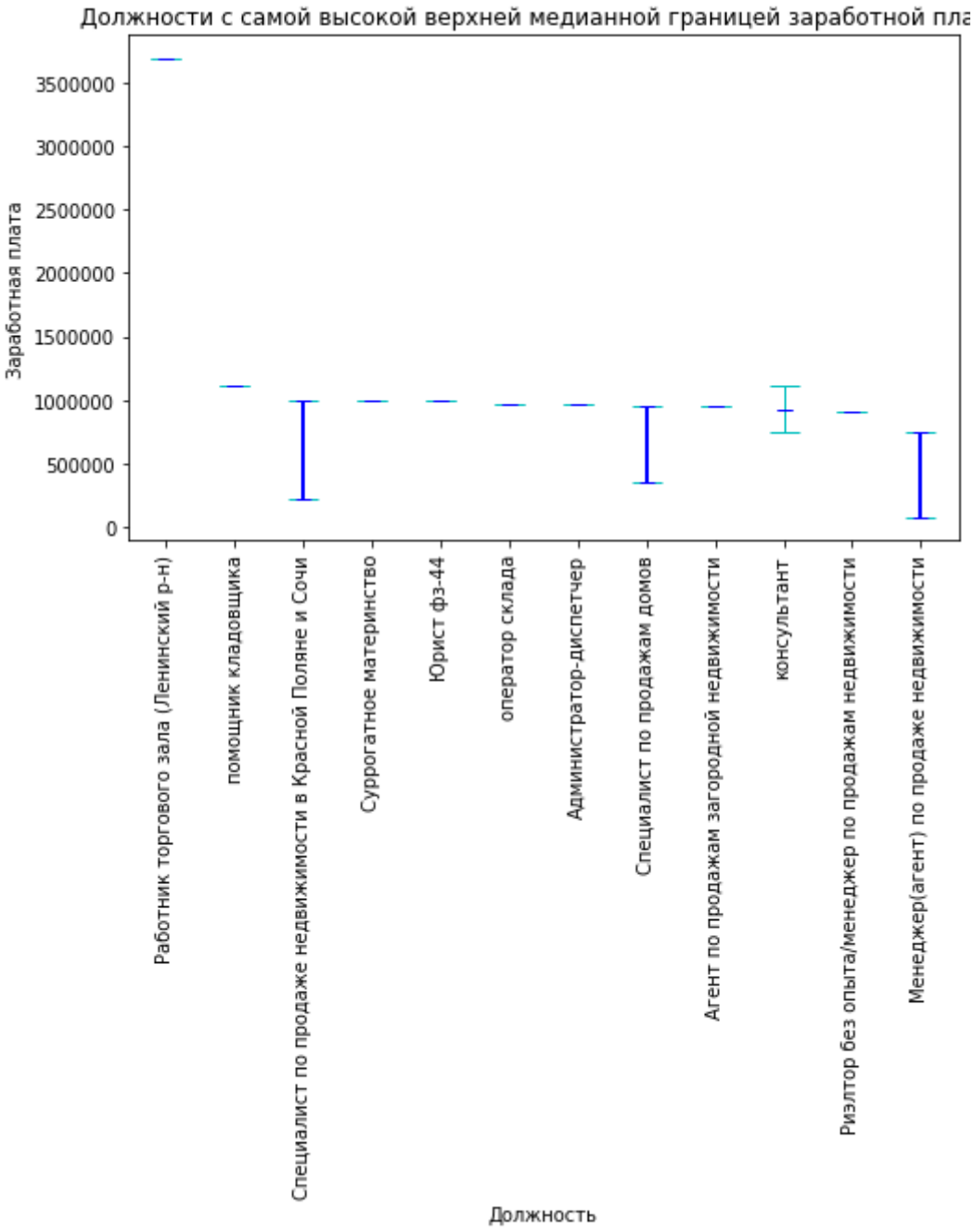

Рисунок 10 – Зависимость медианной верхней границы предлагаемой заработной платы от должности (топ 12)

Среди должностей с наибольшей медианной верхней границей заработной платы много должностей, имеющих также наибольшую абсолютную верхнюю границу. Стоит отметить, что довольно много должностей представлено единичными вакансиями, что объясняет их место в этом топе.

Отдельный интерес представляет «Работник торгового зала (Ленинский р-н)». В качестве нанимателя указан «Работут.», место работы находится в Кемерово. Объяснения такой высокой предлагаемой заработной платы не вижу. Даже если это заработная плата за год, цифра всё равно очень большая.

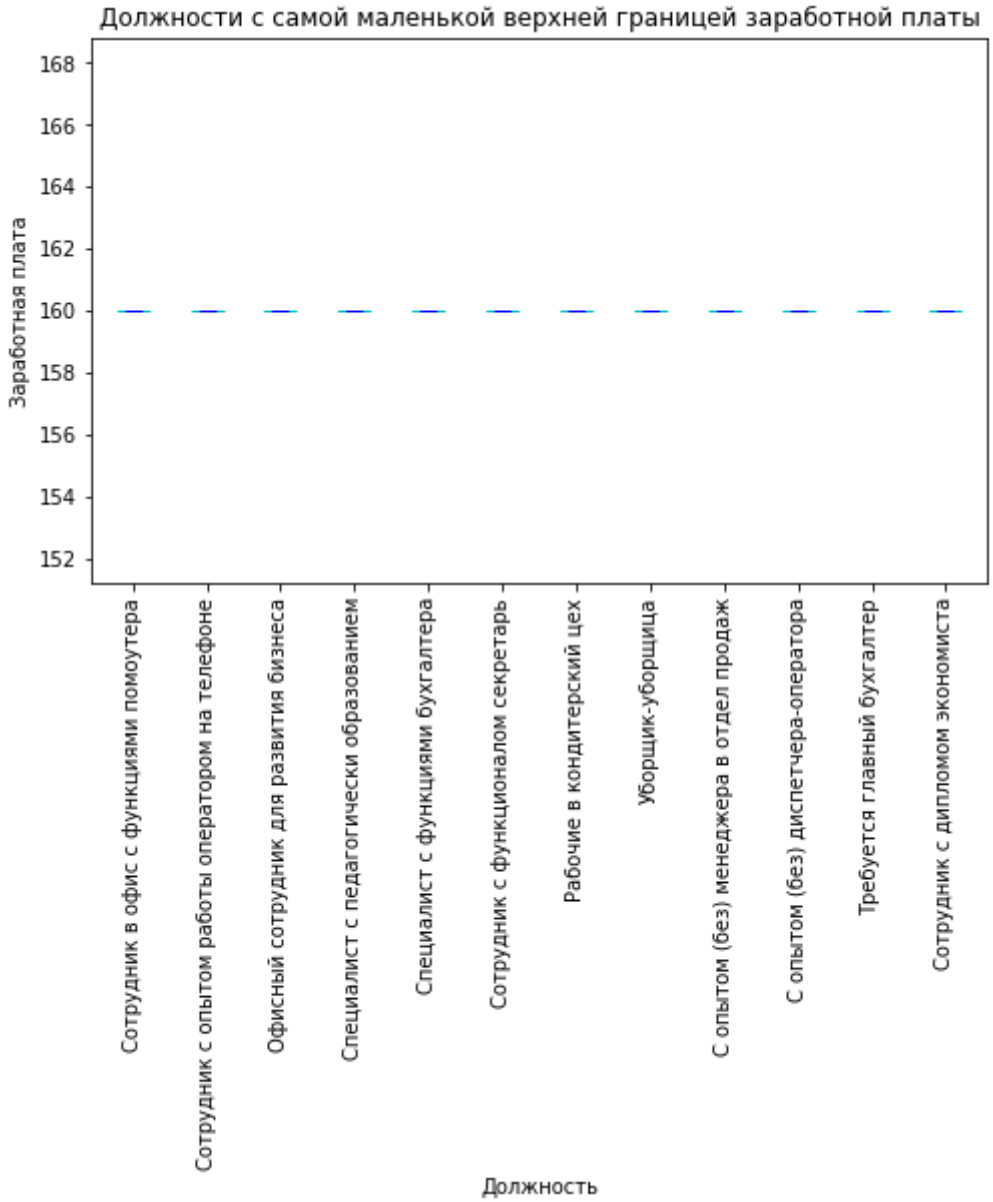

Рисунок 11 – Обратная зависимость абсолютной верхней границы предлагаемой заработной платы от предлагаемой должности (топ 12)

Все вакансии в данном анти-топе имеют одну и ту же предлагаемую заработную плату: 160 ₽. Объяснения столь низкой предлагаемой заработной платы не вижу. Если бы вакансия предполагала 160 тысяч рублей (опустили тысячи), то это было бы слишком много для таких должностей. Если бы у вакансий не была указана валюта, то можно было бы подумать, что подразумевается заработная плата в долларовом эквиваленте, что могло бы выглядеть правдоподобно, но в вакансии явно указана валюта ₽.

Иронично выглядит вакансия «Сотрудник с дипломом экономиста». Кажется, где-то диплом пригодился. Вдвойне иронично видеть такую предлагаемую заработную плату для этой вакансии.

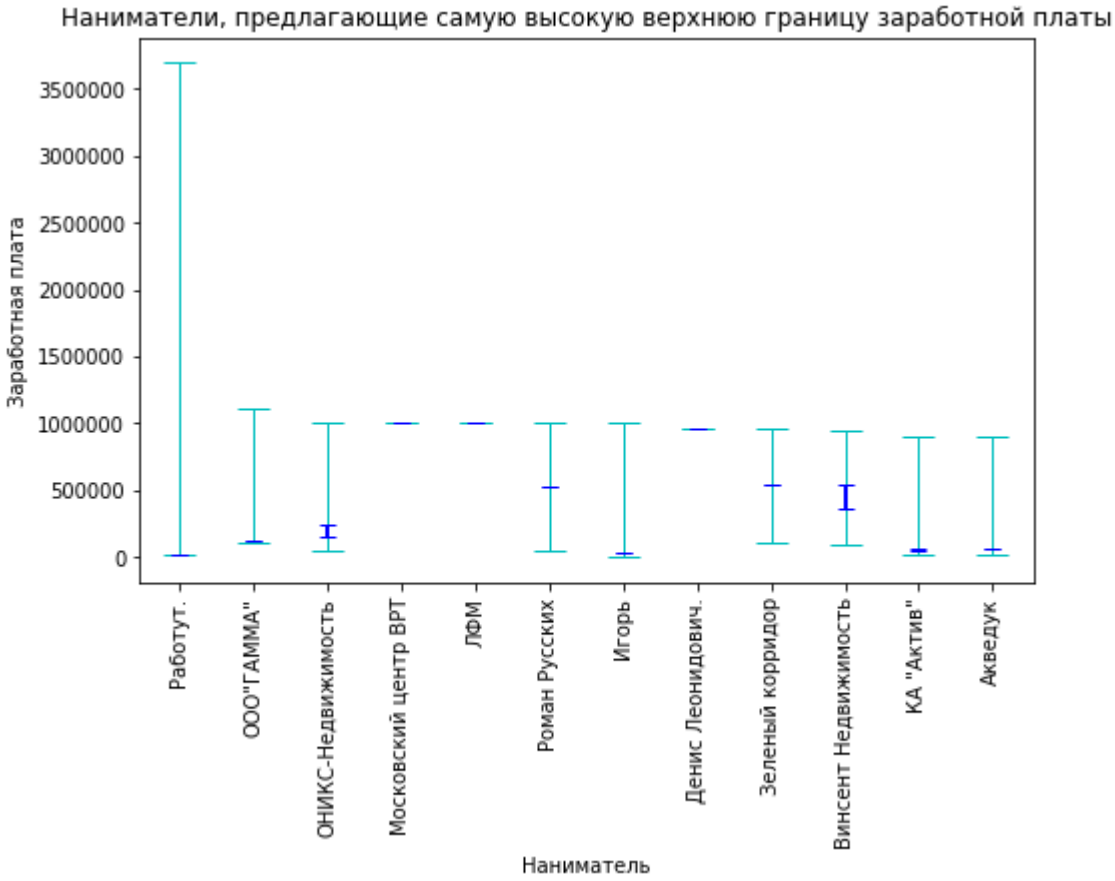

Рисунок 12 – Зависимость абсолютной верхней границы предлагаемой заработной платы от нанимателя (топ 12)

Как и ожидалось, на первом месте фигурирует «Работут.» с вакансией «Работник торгового зала (Ленинский р-н)», за которую предлагается 3696000 ₽. Стоит отметить, что остальные вакансии данного нанимателя выглядят крайне обычно. При такой высокой верхней границе медианная предлагаемая заработная плата составляет всего 20000 ₽.

Также стоит отметить, что в топе фигурируют физические лица и/или ИП.

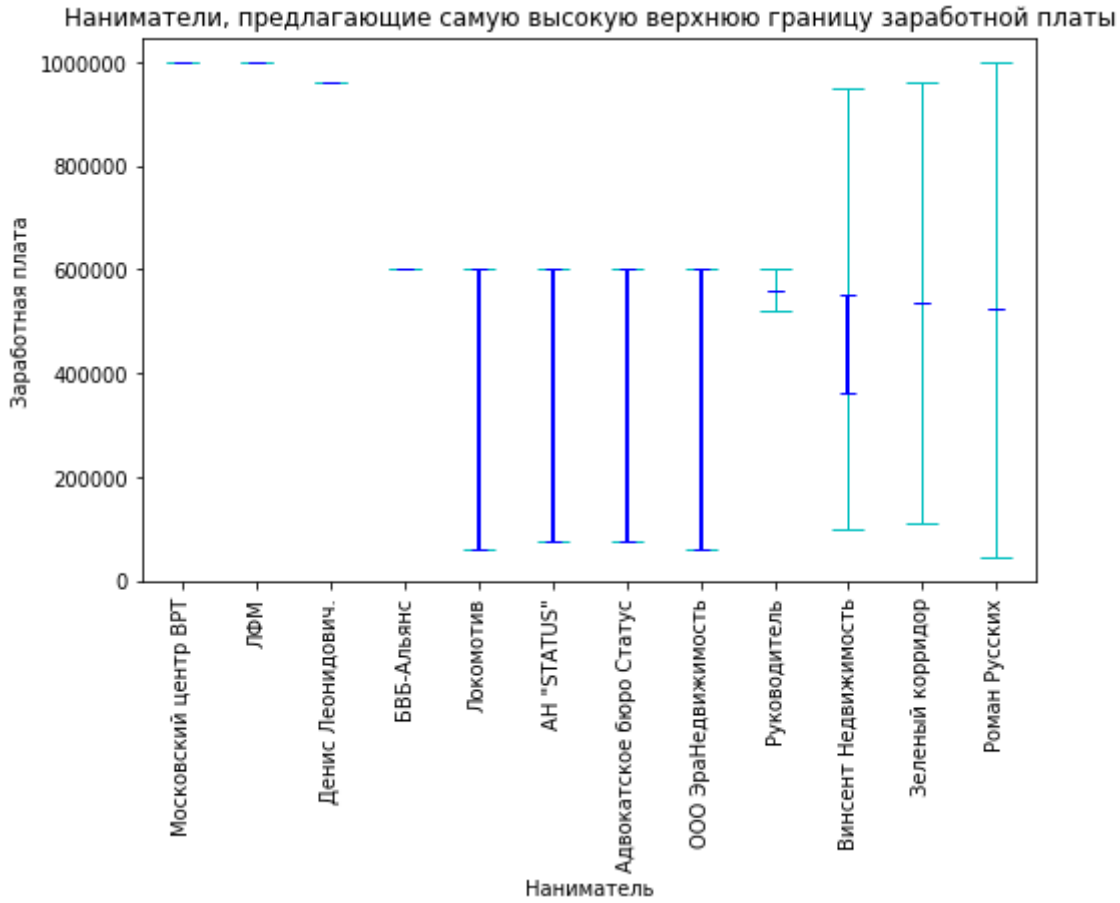

Рисунок 13 – Зависимость медианной верхней границы предлагаемой заработной платы от нанимателя (топ 12)

Среди нанимателей с наибольшей медианной верхней границей заработной платы фигурирует лишь три нанимателя с наибольшей абсолютной верхней границей заработной оплаты, однако они располагаются на трёх первых местах.

Посчитанные абсолютные границы предлагаемой заработной платы: от 160 до 3696000 ₽.

Посчитанные медианные границы предлагаемой заработной платы: от 35000 до 40000 ₽.

Посчитанные средние границы предлагаемой заработной платы: от 44815 до 50260 ₽.

Медианная зарплата по России за апрель 2019-го года составила 34335 ₽, что чуть ниже посчитанных медианных границ. Это объясняется тем, что людям значительно чаще платят ближе к минимальной границе предлагаемой заработной платы, а также тем, что в расчётах учитывались высокие верхние и нижние границы заработной платы, которые, вероятно, используются для привлечения внимание, но не отражают реальную заработную плату.

Средняя зарплата по России составляет 47600 ₽, что очень точно совпадает с полученными данными.

Несмотря на падение российского рубля, по текущему курсу медианная зарплата в России всё ещё значительно (почти в 1.5 раза) превосходит медианную зарплату по Беларуси.

Выводы

Результат выполнения лабораторной работы: при помощи языка программирования Python и фреймворка Scrapy выгружены 204800 вакансии https://russia.trud.com (итоговый файл весил более 100 МБ, выгрузка заняла несколько часов), при помощи стандартной библиотеки языка программирования Python и библиотеки matplotlib выполнены обработка и анализ выгруженных данных, построены графики. Также было выполнено сравнение данных по России с данными по Беларуси, полученными в рамках другой лабораторной работы. Поставленная задача выполнена полностью.

В рамках анализа данных о вакансиях было изучено распределение вакансий по предлагаемым должностям и нанимателям. Исследовались как абсолютные границы заработной платы, так и медианные. Посчитанный медианный уровень предлагаемой в вакансиях заработной платы чуть ниже, чем значение, предоставленное в официальной статистике. При этом сделан вывод, что людям значительно чаще платят ближе к минимальной границе предлагаемой заработной платы, а также, что довольно часто используются высокие границы заработной платы для привлечения внимания. Очевидно, что никто не будет платить 3696000 ₽ работнику торгового зала. При этом

посчитанный средний уровень заработной платы очень точно совпал с официальной статистикой.

Чему научился в рамках выполнения лабораторной работы:

1. Развил навыки выгрузки данных с сайтов с использованием языка программирования Python и фреймворка Scrapy.
2. Изучил и применил на практике способы обхода защиты сайтов от роботов.
3. Развил навыки работы с библиотекой matplotlib, используемой для построения графиков.
4. Развил навыки работы с языком программирования Python. В процессе выполнения лабораторной работы использовалось большое число модулей стандартной библиотеки: re (регулярные выражения), os (функционал ОС), logging (логирование), io (работа с файлами), json (обработка json), collections (структуры данных), statistics (функции математической статистики).