

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.

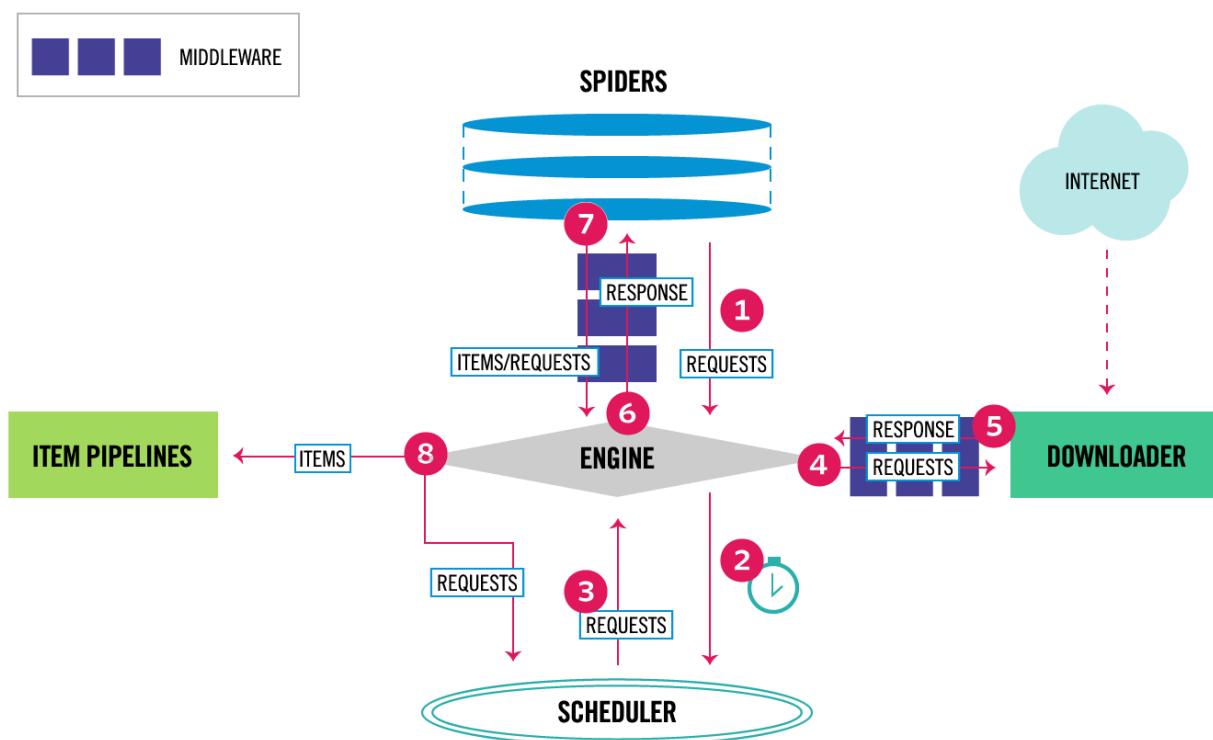


Рисунок 1 – Архитектура фреймворка Scrapy

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.

ГОРОДА:

Работа в Москве

623 140

Работа в Санкт-Петербурге

307 429

Работа в Новосибирске

115 991

Работа в Екатеринбурге

65 821

Работа в Нижнем Новгороде

65 087

Работа в Краснодаре

60 155

Работа в Красноярске

50 253

Работа в Уфе

48 872

Работа в Воронеже

47 535

Работа в Омске

45 716

Работа в Казани

44 572

Работа в Ростове-на-Дону

43 737

Работа в Челябинске

43 529

Работа в Самаре

41 012

Работа в Волгограде

39 350

Работа в Перми

36 699

Работа в Саратове

36 585

Работа в Иркутске

32 657

Работа в Тюмени

32 471

Работа в Ярославле

32 319

Работа в Туле

30 950

Работа в Барнауле

29 065

Работа в Ижевске

27 085

Работа в Твери

26 924

Работа в Хабаровске

26 410

Работа в Рязани

25 144

Работа в Липецке

24 007

Работа в Калуге

23 931

Работа в Ульяновске

23 709

Работа в Белгороде

22 652

Работа в Кемерово

22 461

Работа в Томске

22 342

Работа в Владимире

21 796

Работа в Калининграде

20 593

Работа в Тольятти

20 055

Работа в Набережных Челнах

19 863

Работа в Владивостоке

19 586

Работа в Пензе

19 529

Работа в Иваново

19 181

Работа в Новокузнецке

18 881

Работа в Оренбурге

18 739

Работа в Курске

17 791

Работа в Брянске

17 480

Работа в Сочи

17 284

Работа в Чебоксарах

17 093

Работа в Кирове

16 825

Работа в Смоленске

16 386

Работа в Мурманске

16 351

Работа в Архангельске

16 117

Работа в Костроме

16 108

транспортная компания "Автовозим" на полную занятость, в офисе, зарплата на руки

Агент по недвижимости (купля и продажа)

Москва

• МИЭЛЬ

★★★★★

2 отзыва

В компанию Мизель на полный рабочий день требуется Агент по недвижимости (купля и продажа), график работы 5/2, один год. Зарплата на руки по договорённости. Работа в

Водитель вилочного погрузчика

40000-45000 Р

Ханты-Мансийский автономный округ, Сургут • ООО "Меланж"

Нужен Водитель вилочного погрузчика в компанию ООО "Меланж" на полную занятость, в офисе. Зарплата на руки до 45000. . 40 часов в неделю. Требуемый опыт

Продавец-кассир

30000-40000 Р

Московская область, Истра • Дмитрогорский продукт

Нужен Продавец-кассир в компанию Дмитрогорский продукт на полную занятость, в магазине. Зарплата на руки до 40000. Продавец-кассир приветствует и консультирует

Менеджер выездных продаж

40000 Р

Свердловская область, Нижний Тагил • СКБ Контур

★★★★★

2 отзыва

В компанию СКБ Контур на полный рабочий день требуется Менеджер выездных продаж, график работы 5/2, без опыта. Зарплата на руки от 40000. Работа в офисе.

Водитель на миксер / АБС

70000 Р

Москва • Группа компаний Веста-СФ

Доставка бетона и раствора на строительные объекты по г. Москве и ближайшему Подмосковию.- Опыт на миксере (база МАЗ, КАМАЗ). ПРОСЬБА - ПО ВАКАНСИИ -

Инженер-проектировщик НВК (водопровод, канализация, ...

Москва, Кузьминки • Группа компаний Веста-СФ

- выполнение проектной и рабочей документации по наружным сетям Водоснабжения и Водоотведения; - выполнение гидравлических расчетов напорных и безнапорных

Агент по работе с недвижимостью (вторичный рынок жиль...

60000 Р

Москва • ИНКОМ НЕДВИЖИМОСТЬ

★★★★★

9 отзывов

Обязанности: - Проведение переговоров - Заключение договоров купли - продажи объектов недвижимости - Сопровождение сделки - Оформление документов

Укладчик тротуарной плитки

50000 Р

Москва • Группа компаний Веста-СФ

Укладка тротуарной плитки, брусчатки, установка бордюрного камня, устройство

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.

4

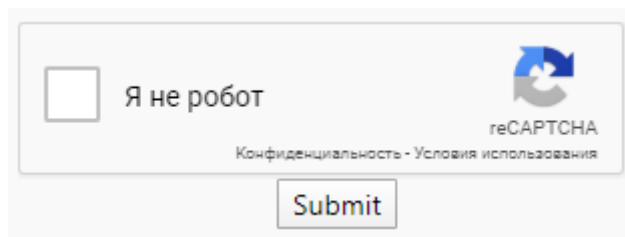


Рисунок 3 – Защита от роботов

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 Р", "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 ... "}.

```
1 [{"name": "Требуется: строительно-отделочные бригады для выполнения работ по Свердловской области", "salary": "6000 Р", "place": "Москва", "employer": "Ми  
2 [{"name": "Преподаватель курса \"Развитие памяти и скорочтения\"", "salary": "15000-30000 Р", "place": "Москва", "employer": "Экзимум", "description": "Опи  
3 [{"name": "WEB-программист", "salary": "200000 Р", "place": "Москва", "employer": "Работа-вакансии.рф", "description": "Создание и поддержка веб-сервиса.  
4 [{"name": "Требуется: Семейный водитель, Переделкино.", "salary": null, "place": "Москва", "employer": "Империя", "description": "Обязанности: Вождение а  
5 [{"name": "Senior Business analyst", "salary": null, "place": "Moscow, Russian Federation", "employer": "Luxoft", "description": "Project-Description Зак  
6 [{"name": "Помошник системного администратора", "salary": "60000 Р", "place": "Москва", "employer": null, "description": "Компания предлагает работу помо  
7 [{"name": "Пекарь", "salary": "37000-40000 Р", "place": "Москва", "employer": "Пятерочка", "description": "Описание вакансии: \\\nДля каждого, кто хочет ра  
8 [{"name": "DevOps-инженер", "salary": null, "place": "Москва", "employer": "ЕКС", "description": "В Мир Инвестиций\ БКС мы ищем DevOps-инженера. Мы пре  
9 [{"name": "Программист, на дому", "salary": null, "place": "Москва", "employer": "Центр оценки квалификаций Совета по проф.квалификациям финансового рынка  
10 [{"name": "Мойщик-уборщик подвижного состава", "salary": null, "place": "Москва", "employer": "ГУП Мосгортранс филиал Служба материально-технического об  
11 [{"name": "Инженер по эксплуатации и ремонту (климатическое оборудование)", "salary": null, "place": "Moscow, Russian Federation", "employer": "Simens",  
12 [{"name": "Упаковщик на склад мясокомбината", "salary": null, "place": "Москва", "employer": "Энергошит", "description": "Все наши объекты являются жизне  
13 [{"name": "Менеджер", "salary": "30000 Р", "place": "Москва, Московская область и Москва", "employer": null, "description": "Регистрируйтесь в отличном,  
14 [{"name": "Охранник (вахта)", "salary": "45000 Р", "place": "Москва, метро Аннино, Аннино", "employer": "Инконсалт К", "description": "Охранник / Сотрудни  
15 [{"name": "Генеральный директор акционерного общества", "salary": null, "place": "Москва", "employer": null, "description": "Требования: Профессиональное  
16 [{"name": "Специалисты в онлайн-проект", "salary": "25000 Р", "place": "Москва", "employer": "ООО Северина", "description": "В онлайн-проект требуется к  
17 [{"name": "Оператор проекционной аппаратуры и газорезательных машин", "salary": "75000 Р", "place": "Отрадное", "employer": null, "description": "Информ  
18 [{"name": "Швея", "salary": "50000 Р", "place": "Москва", "employer": "ООО PAIVITA", "description": "Должность: ШвеяОписание компании: Компания Dimanche L  
19 [{"name": "Юрисконсульт", "salary": null, "place": "Москва", "employer": "MERLION", "description": "MERLION сегодня - это 12 000 профессионалов, которые  
20 [{"name": "Визуальный мерчендайзер (м. Павелецкая)", "salary": "50000 Р", "place": "Москва, м. Павелецкая", "employer": "Модное Сбро", "description": "Усл  
21 [{"name": "Разнорабочий", "salary": "40000 Р", "place": "Москва, Деловск, Ногинск, Одинцово, Рошаль, м. Кра...", "employer": "ИП \"Рыарева Н.В.\", "descri  
22 [{"name": "Комплектовщик", "salary": "33000-43000 Р", "place": "Москва, г. Москва", "employer": "Инконсалт К", "description": "Обязанности: Упаковка на с  
23 [{"name": "Работник торгового зала", "salary": null, "place": "Москва", "employer": "микс ки, ООО", "description": "Требования: до 45 летОбязанности:Обяз  
24 [{"name": "КОМПЛЕКТОВЩИК (-ЦА)", "salary": null, "place": "г. Москва Метро КРАСНОГВАРДЕЙСКАЯ, ВАРШАВСКАЯ, НАГ...", "employer": null, "description": "--Т  
25 [{"name": "Полицейский", "salary": "44000-55000 Р", "place": "Москва, Сигнальный проезд, Ас", "employer": "Полк охраны и конвоирования подозреваемых и обви  
26 [{"name": "Продавец (м. Смоленская)", "salary": "40000-50000 Р", "place": "Москва, Библиотека им. Ленина", "employer": "Азбука вкуса", "description": "П  
27 [{"name": "Фасовщик в тёплый склад продуктов питания", "salary": "43500-54100 Р", "place": "Москва", "employer": null, "description": "Звоните! Можно Wha  
28 [{"name": "Java-разработчик", "salary": "300000 Р", "place": "Москва", "employer": "Почта России", "description": "Технологии цифровых данных проникает в  
29 [{"name": "Грузчик-разнорабочий-теплицы (вахта)", "salary": "62400 Р", "place": "Москва", "employer": "Лайн-логистик", "description": "РАБОТАМ В ОДНОМ  
30 [{"name": "Эксперт-преподаватель образовательных программ по направлению Cyber Security", "salary": null, "place": "Москва", "employer": "ООО Хакер", "de  
31 [{"name": "Грузчик на рид (вахта)", "salary": "30000 Р", "place": "Москва", "employer": "Амстафф", "description": "Растружка и погрузка готовой про  
32 [{"name": "Помошник менеджера по организации мероприятий", "salary": null, "place": "Москва", "employer": "Бизнес-Ассистанс", "description": "Описание: П  
33 [{"name": "Зоо-специалист", "salary": null, "place": "Москва, Московская область", "employer": "Кристалл", "description": "Стажировка полностью дистанция  
34 [{"name": "Охранник в офис", "salary": "45000 Р", "place": "Москва, метро Пушкинская, Тверская", "employer": "Инконсалт К", "description": "Охранник / со  
35 [{"name": "Комплектовщик-фасовщик вахта", "salary": "42800-76000 Р", "place": "Москва, метро Ясенево, Ясенево", "employer": "ПРЕМЬЕР-ИНСАЙТИНГ", "desc  
36 [{"name": "Сборщики накладных", "salary": "36000-75000 Р", "place": "Москва", "employer": "ООО \"Бид групп\"", "description": "В интернет-магазин требуют  
37 [{"name": "Водитель категории В, на автомобиле компании", "salary": "30000 Р", "place": "Москва, Пушкинская", "employer": "Центр Найма Водителей", "descri  
38 [{"name": "Кладовщик-приемщик", "salary": "38600-62000 Р", "place": "Москва, Московская обл. Раменский р-н", "employer": "ООО \"Тендерные технологии\"", "de  
39 [{"name": "Грузчик спортивной одежды", "salary": "26400 Р", "place": "Москва", "employer": "Энергошит", "description": "Вахта с проживанием, Москва, МО.  
40 [{"name": "Dev/Full-developer", "salary": null, "place": "Москва, Долгоруковский проезд", "employer": "Аэросис-Иллюминация", "description": "Вакансия с ра  
41 [{"name": "Специалист по подбору персонала", "salary": null, "place": "Москва, Долгоруковский проезд", "employer": "Аэросис-Иллюминация", "description": "Вакансия с ра
```

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 Р). 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**

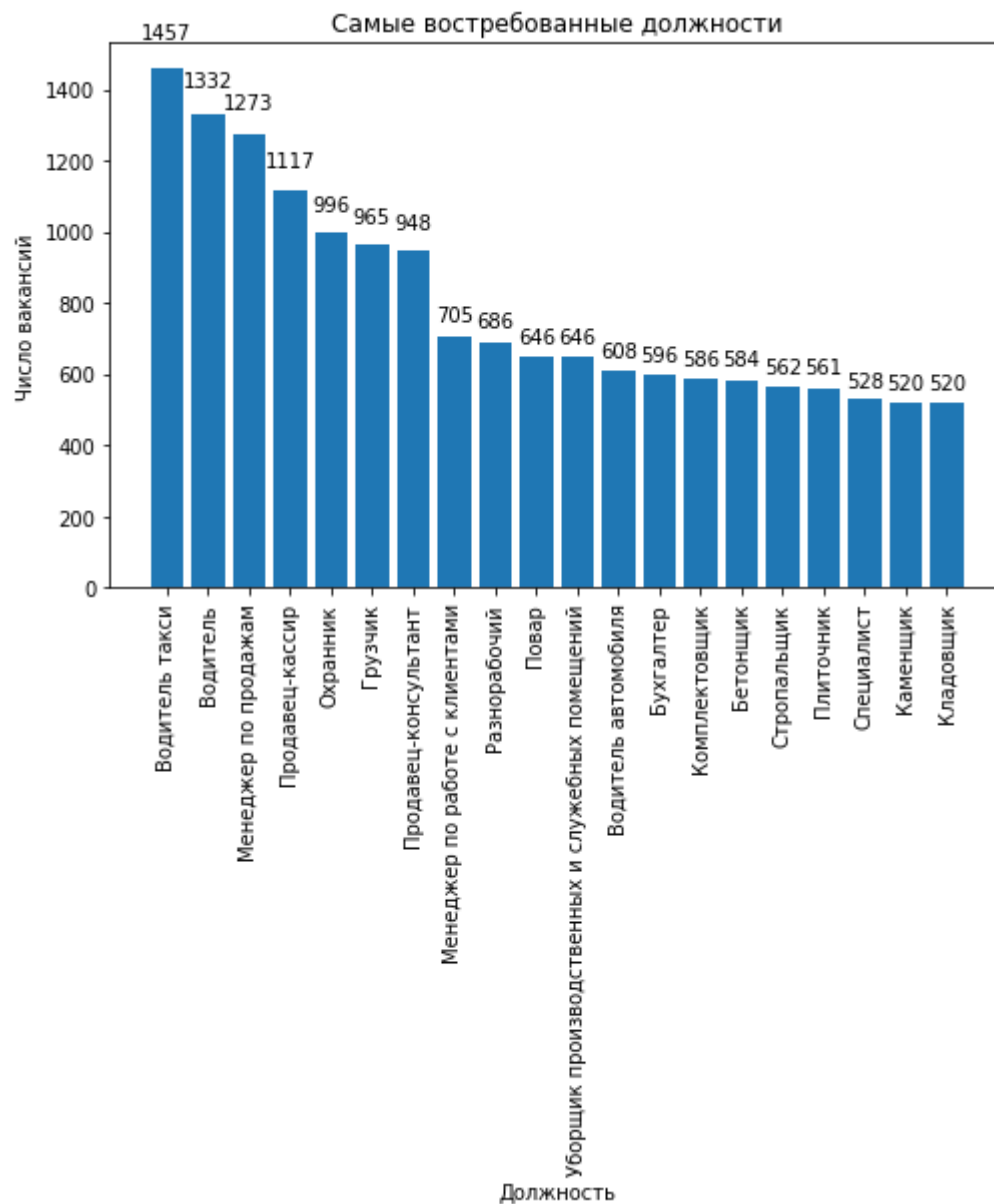


Рисунок 5 – Распределение вакансий по предлагаемым должностям (топ 20 должностей)

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.

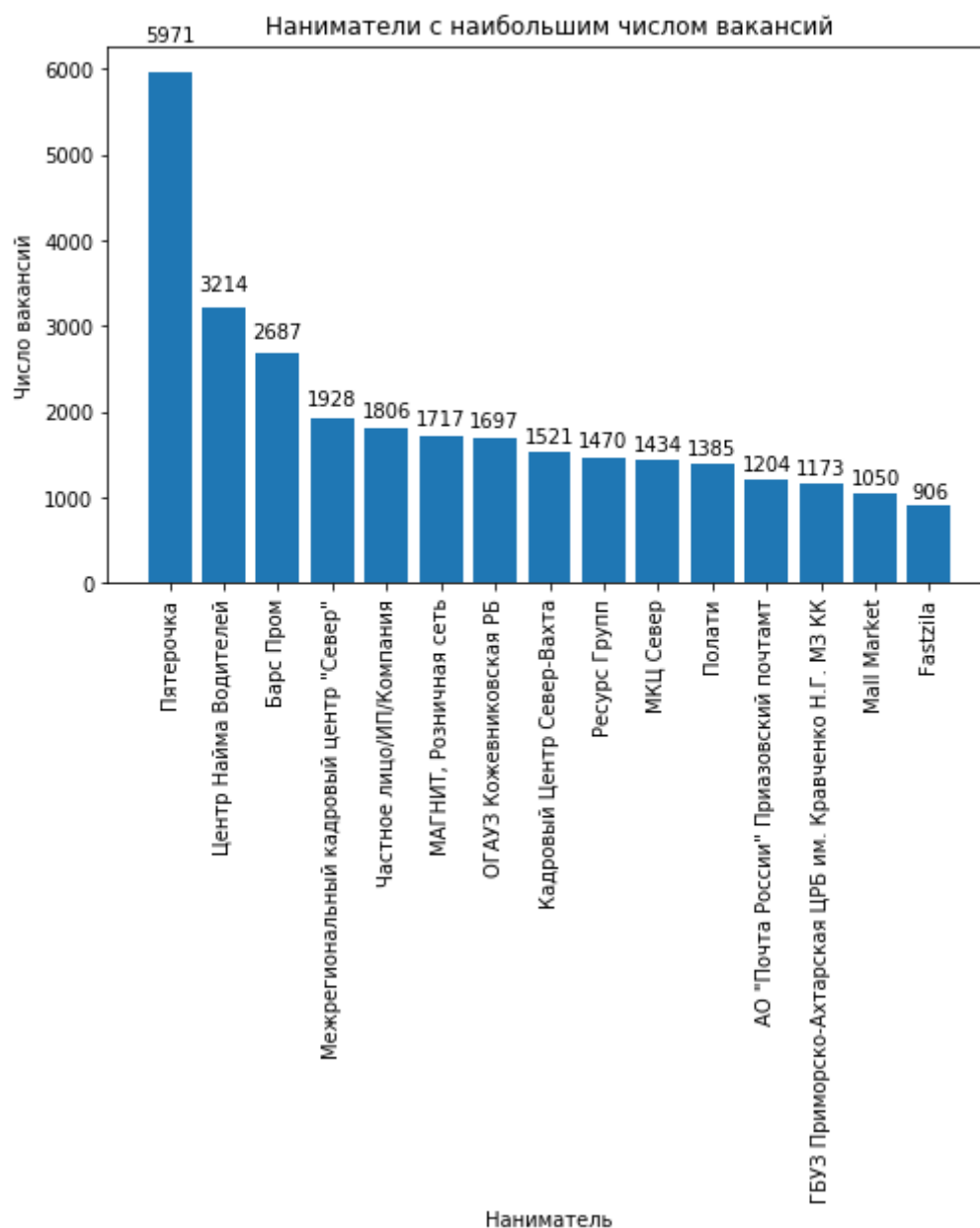


Рисунок 6 – Распределение вакансий по нанимателям (топ 20 значений)

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.

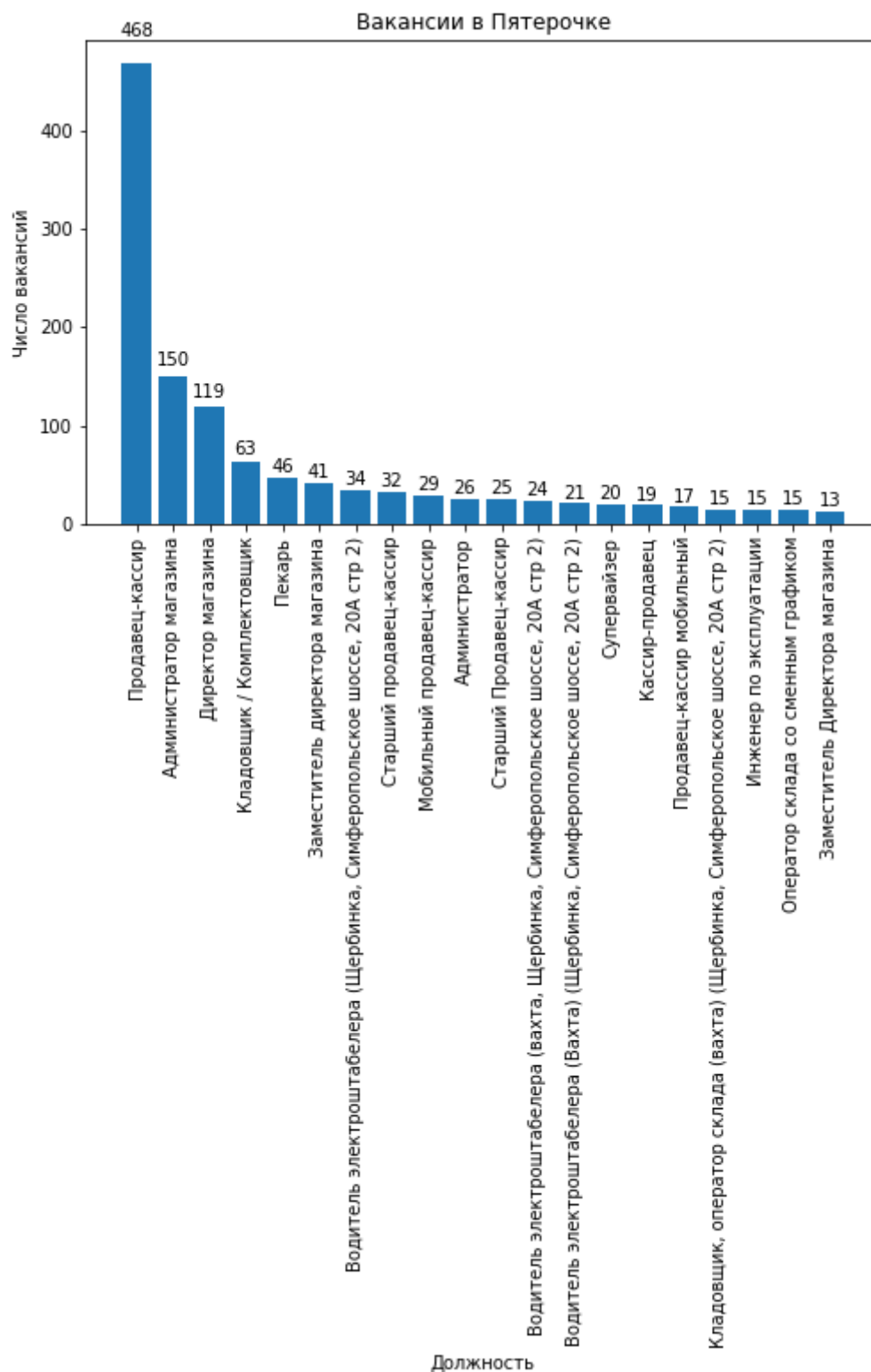


Рисунок 7 – Распределение вакансий по предлагаемым должностям в «Пятерочке» (топ 20 должностей)

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

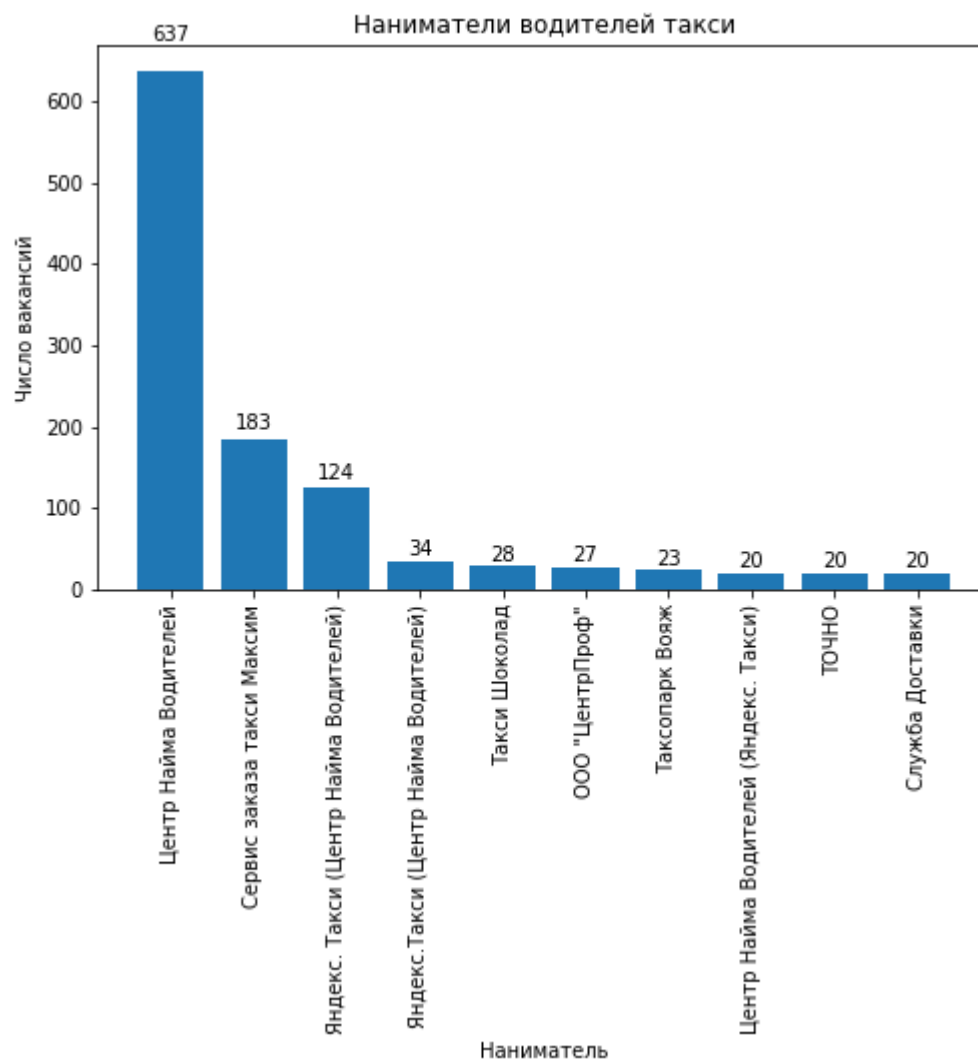


Рисунок 8 – Распределение вакансий «водитель такси» по нанимателям (топ 10 значений)

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

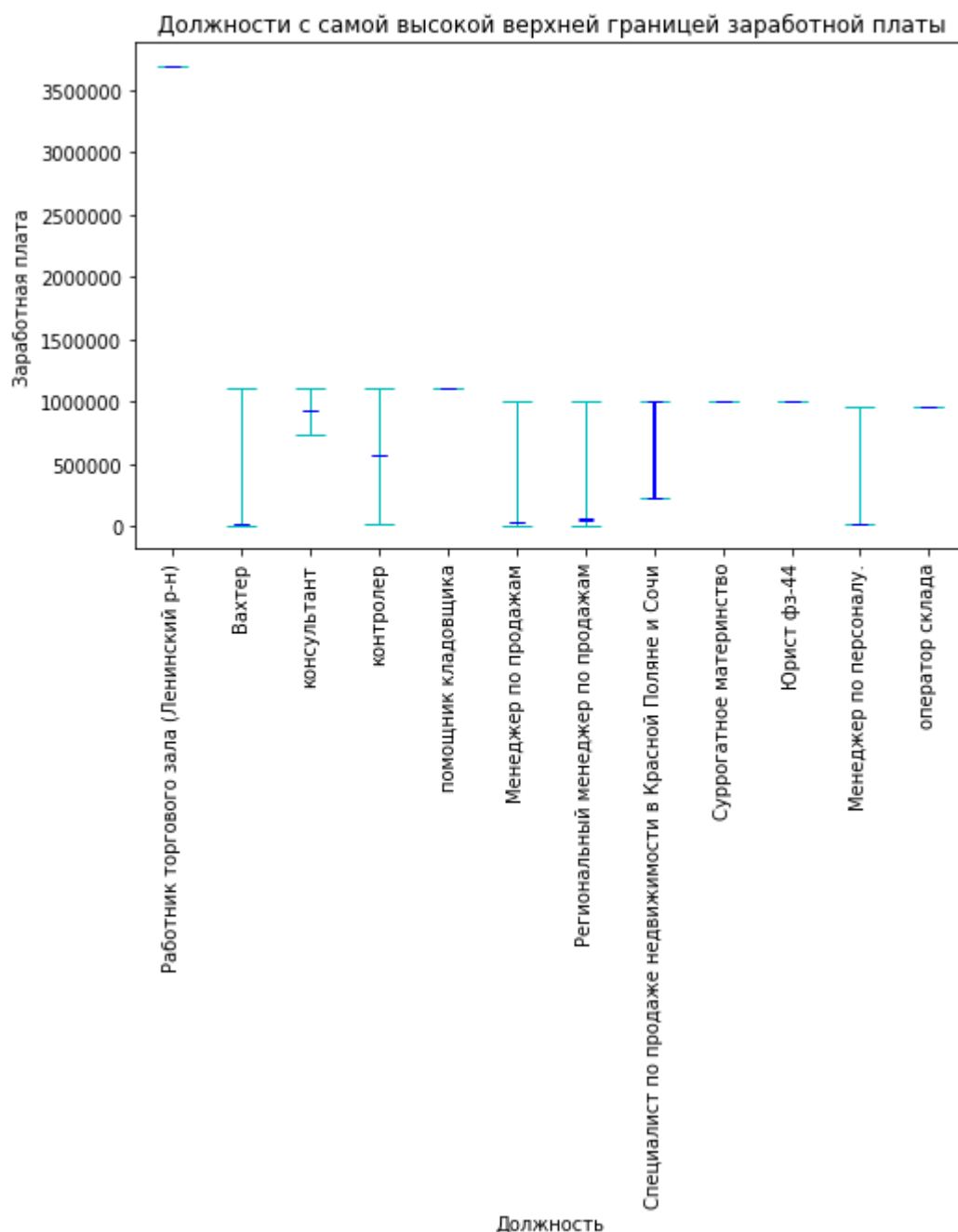


Рисунок 9 – Зависимость абсолютной верхней границы предлагаемой заработной платы от должности (топ 12)

Интересно видеть, что должности с самой высокой возможной заработной платой также имеют очень низкую нижнюю границу. Зарботная плата вакансии "Вахтер" может варьироваться от 9500 до 1109880 Р, 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.

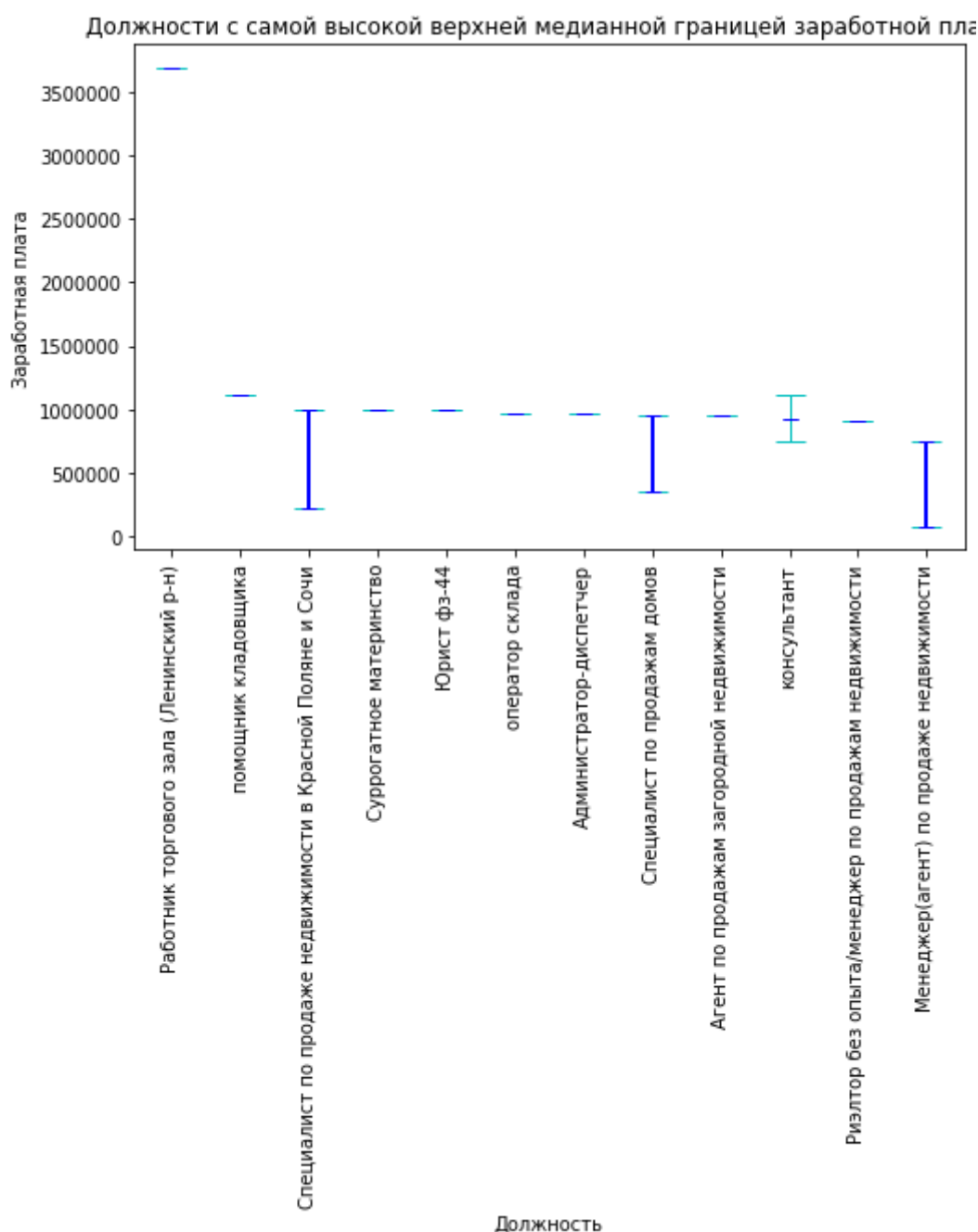


Рисунок 10 – Зависимость медианной верхней границы предлагаемой заработной платы от должности (топ 12)

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.

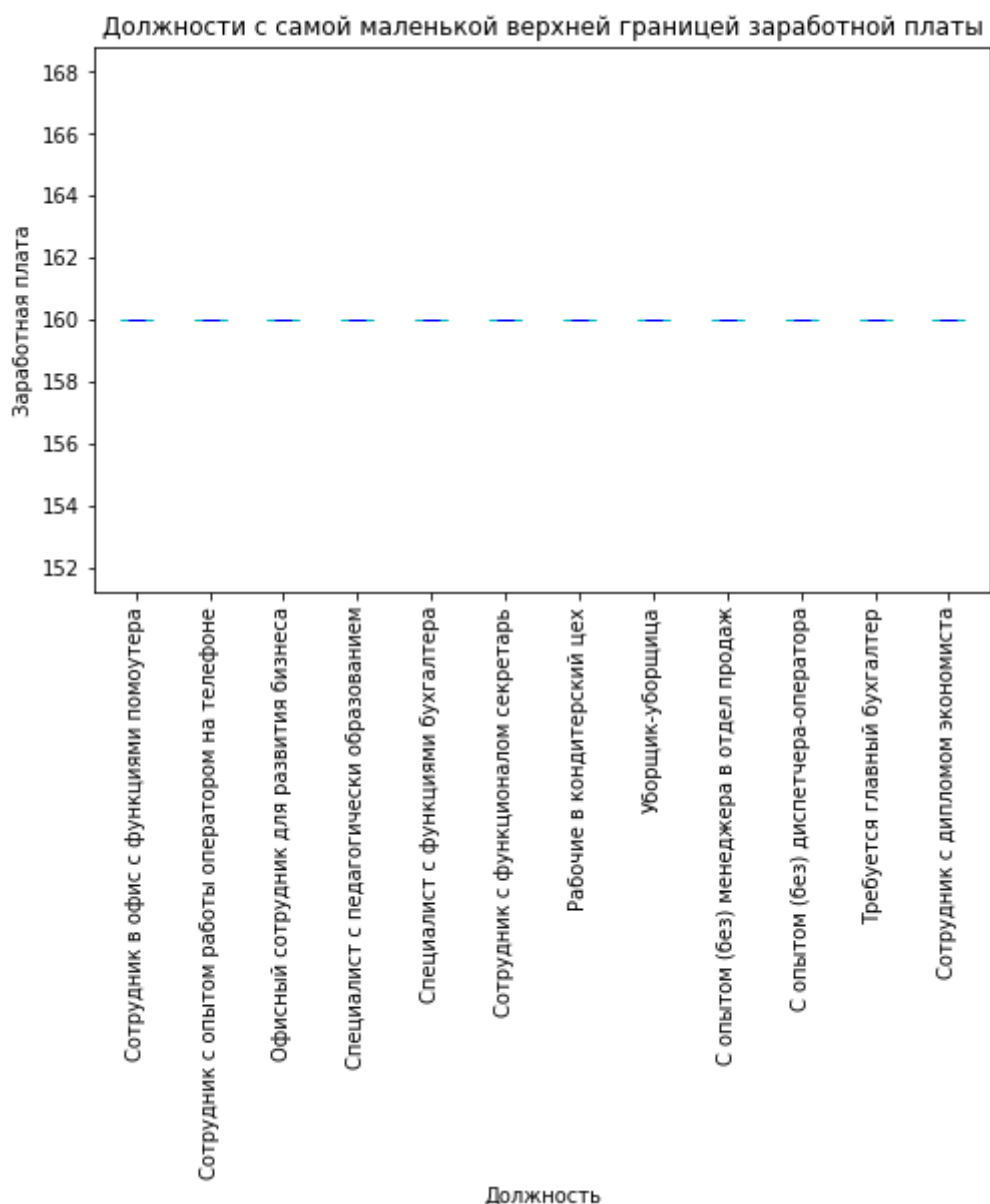


Рисунок 11 – Обратная зависимость абсолютной верхней границы предлагаемой заработной платы от предлагаемой должности (топ 12)

Все вакансии в данном анти-топе имеют одну и ту же предлагаемую заработную плату: 160 Р. Объяснения столь низкой предлагаемой заработной платы не вижу. Если бы вакансии 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 Р.

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

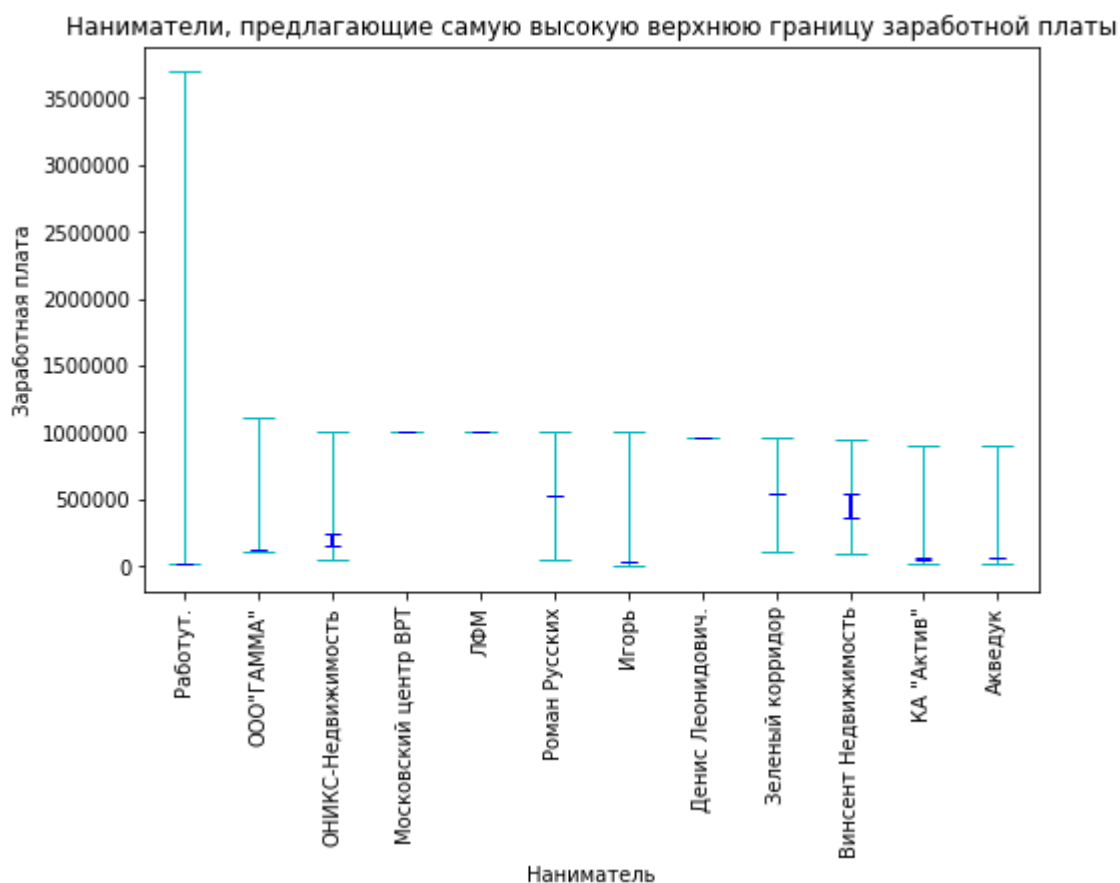


Рисунок 12 – Зависимость абсолютной верхней границы предлагаемой заработной платы от нанимателя (топ 12)

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

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

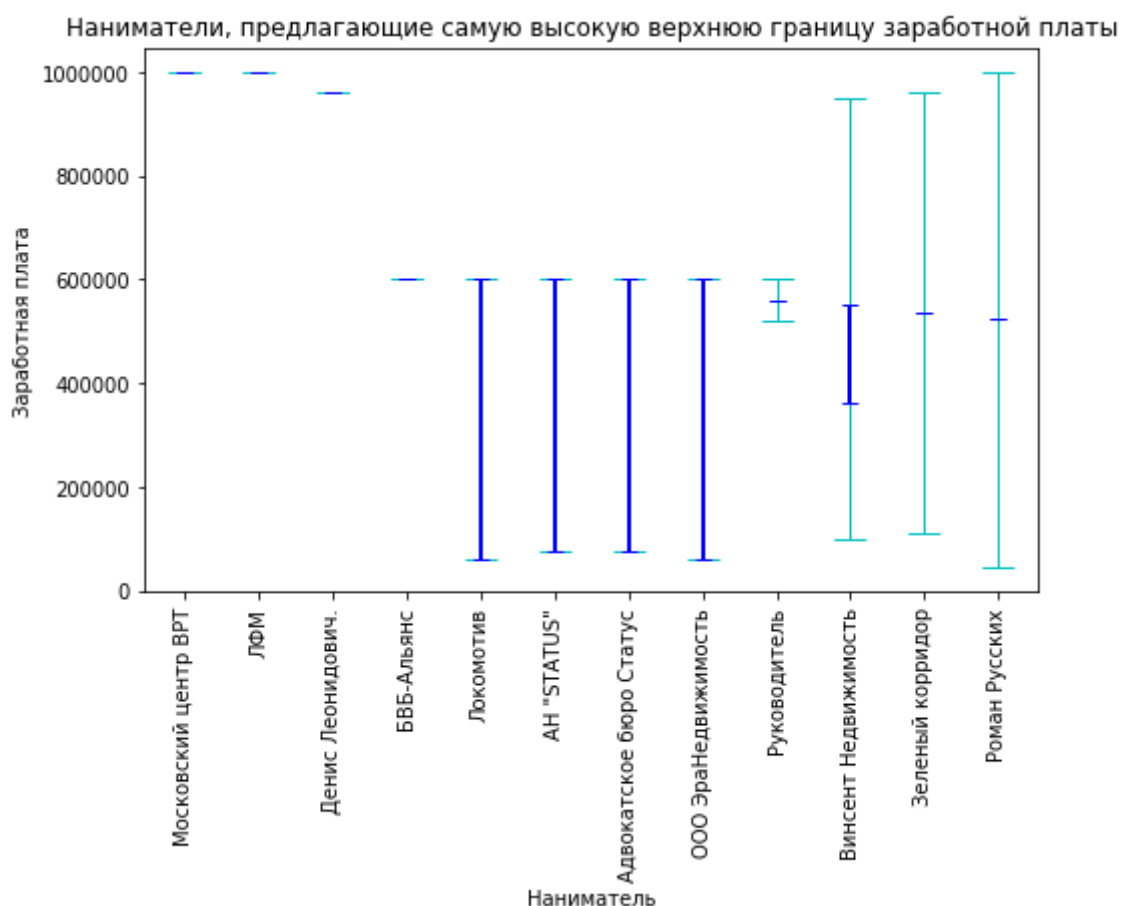


Рисунок 13 – Зависимость медианной верхней границы предлагаемой заработной платы от нанимателя (топ 12)

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

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

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

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

Медианная зарплата по России за апрель 2019-го года составила 34335 Р, 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 внимания, но не отражают реальную заработную плату.

Средняя зарплата по России составляет 47600 Р, 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 Р 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).