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 «Модели и методы обработки больших объемов информации»
на тему «**Обработка данных о вакансиях**»

Выполнил студент группы 956241: Зязюлькин С.П.

Проверил: Стержанов М.В.

Минск, 2019

Постановка задачи

Задача заключается в анализе информации о вакансиях в Беларуси. В частности, предполагается изучить распределение вакансий по требуемому образованию, городам, категориям, предлагаемым должностям, нанимателям, дате размещения, предполагается исследовать требуемый для вакансий опыт и размер предлагаемой заработной платы. Также предполагается сравнить медианный уровень предлагаемой в вакансиях заработной платы с официальной статистикой, предоставляемой Национальным статистическим комитетом Беларуси.

План выполнения работы

1. Найти источник данных о вакансиях в Беларуси.
2. Реализовать выгрузку данных из источника.
3. Выполнить обработку загруженных данных.
4. Провести анализ обработанных данных.
5. Сделать выводы.

Получение данных

В качестве источника данных о вакансиях в Беларуси выбран сайт https://rdw.by/vakansii. На момент выгрузку сайт содержал 5390 вакансий. Для каждой вакансии выгружаются следующие данные:

- Дата подачи вакансии.
- Предлагаемая должность.
- Наниматель.
- Предлагаемый уровень заработной платы.
- Требуемый уровень образования.
- Требуемый опыт.
- Город.
- Категория 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.

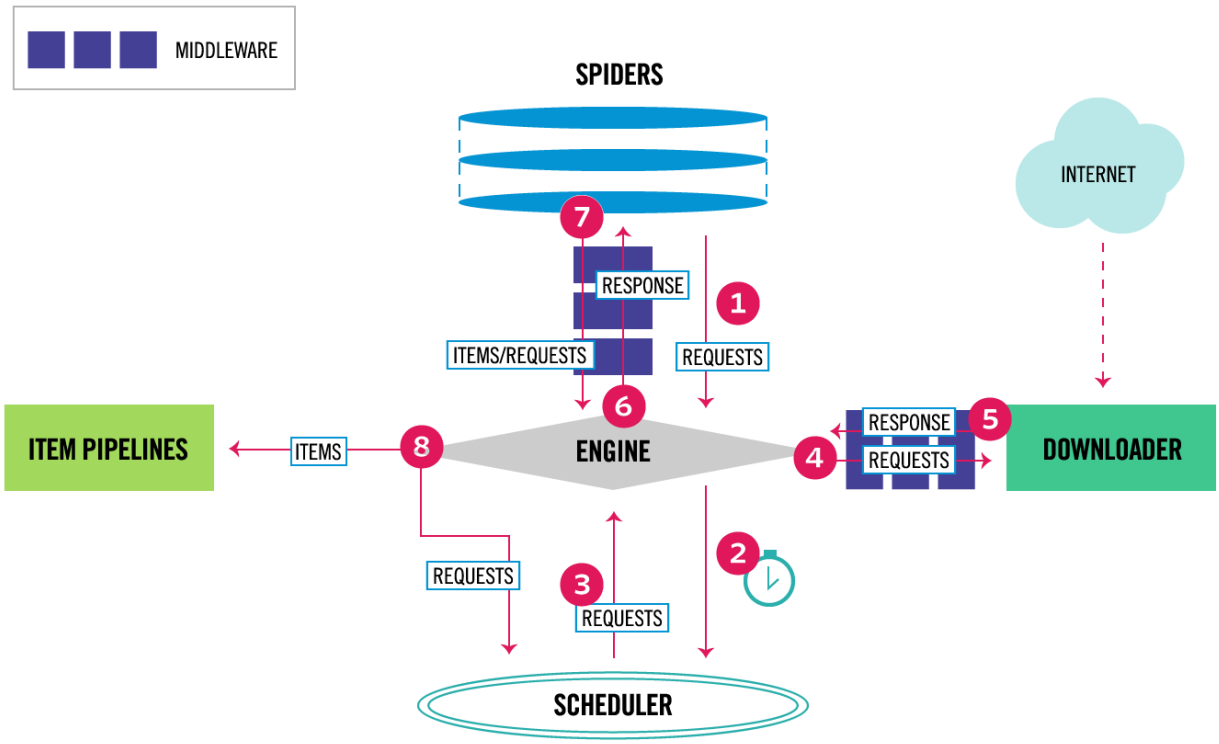

Рисунок 1 – Архитектура фреймворка Scrapy

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

Поиск нужных данных в html-странице выполнялся при помощи css-селекторов.

CSS Selectors

Selector	Role
p{ }	Tag selector, **all p tags**
#para{ }	Id para (***unique***)
.para1{ }	Class para1 (***multiple***)
p.para{ }	P tag with class para
P .para{}	P with child having class para
div p{}	p tag having parent div.
*{}	All tags{ **Universal Selector**}
h1, h3, h5{}	Only h1, h3 and h5 (***grouping***)
.para a{ }	A with parent para class
body{}	Parent of all tags

Рисунок 2 – Основные css-селекторы

Выгруженные данные сохраняются в файл в формате json. Пример данных одной вакансии в формате json: {"date": "30.11.2019 | 09:27", "name": "Водитель кат. С", "employer": "Государственное предприятие «Октябрьский ГК»", "salary": null, "education": "любое", "experience": "от 1 года", "place": "Минск", "category": "Общепит, рестораны, гостиницы, игорный бизнес"}.

Обработка выгруженных данных

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

Действия, выполняемые во время обработки:

1. Из поля "дата подачи вакансии" при помощи регулярного выражения извлекается дата.
2. Если в вакансии отсутствует предлагаемая должность, то она заменяется на "Не указано". Также удаляются специальные символы из имени вакансии.
3. 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: от 1 000 Br, от 100 до 10 000 Br). Одна или обе границы могут быть не указаны.
4. Из поля "требуемый уровень образования" при помощи регулярного выражения удаляются лишние пробельные символы и лишний текст.
5. Из поля "требуемый опыт" при помощи регулярных выражений удаляются лишние пробельные символы и лишний текст, а затем извлекается значение требуемого опыта (пример формата до извлечения значения: от 1 года).
6. Если не указана категория вакансии, то она заменяется на "Не указана".
7. Для всех полей, содержащих категориальные данные, формируется множество допустимых значений. К таким полям относятся следующие: предлагаемая должность, наниматель, требуемый уровень образования, город, категория вакансии.

Категориальные данные

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

- Предлагаемая должность (10 первых значений): 'Преподаватель ландшафтного дизайна', 'Комплектовщик', 'Маляр по металлу', 'Преподаватель курса «дизайн интерьеров»', 'Продавец-консультант в салон обоев', 'Маляр по окраске металлоконструкций', 'Повар-универсал', 'Главный архитектор', 'Инженер-программист 1С', 'Пеший курьер'.
- Наниматель (10 первых значений): 'Представительство «ЕвроАвтоТранс»', 'ПК «Деньги в долг финанс»', 'ЧУП «Лагентранс»', '«Автолайтэкспресс»', 'ООО «Асплант Евро Кемикалс»', 'ООО «СТРИТ ТАКСИ»', 'СЗАО «Лебортово»', ' «Инжмашлогист»', 'ООО «ГолдДекор»', 'ООО «ФСБК-СтройИндустрия»'.
- Требуемый уровень образования: 'среднее', 'профессионально-техническое', 'любое', 'среднее специальное', 'без образования', 'высшее', 'неоконченное высшее'.
- Город (20 первых значений): 'Глубокое', 'Крупки', 'Миоры', 'Гомель', 'Пуховичи', 'Червень', 'Волковыск', 'Бобруйск', 'Раков', 'Кобрин', 'Лоев', 'Орша', 'Березино', 'Слоним', 'Могилев', 'Жлобин', 'Давид-Городок', 'Фаниполь', 'Сморгонь', 'Солигорск'.
- Категория вакансии (5 первых значений): 'Транспорт, автобизнес, автосервис', 'Сельское хозяйство, агробизнес', 'Бытовые услуги, ЖКХ, услуги для населения', 'Образование, наука, культура', 'Не указана'.

• Анализ данных

Анализ данных выполнялся на языке программирования Python с использованием стандартной библиотеки. Для построения графиков использовалась библиотека matplotlib.

На графиках заработной платы будут отображены максимальные и медианные значения границ. Медианные значения границ не столь показательны, т.к. существует большой разброс в количестве данных в категориях. Так, категория с всего несколькими вакансиями, но вакансиями, имеющими высокие показатели заработной платы, будет несправедливо вырываться вперёд. Поэтому сортировка на таких графиках идёт по максимальной верхней границе заработной платы.

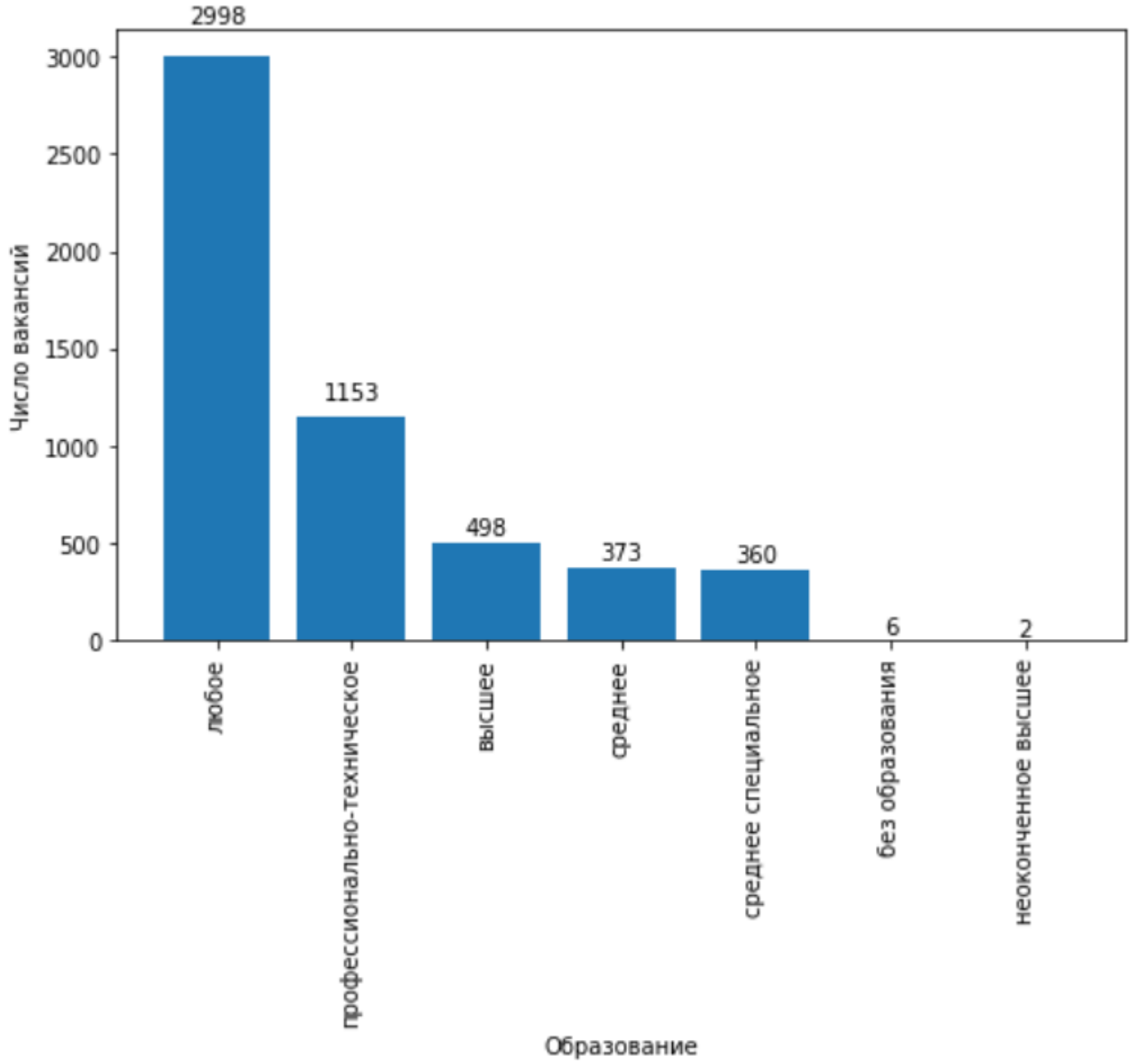

Рисунок 3 – Распределение вакансий по требуемому уровню образования

Больше половины вакансий не требует наличия образования. Сравнительно большое число вакансий требует профессионально-техническое образование. Высшее образование требуется всего лишь для 498 вакансий.

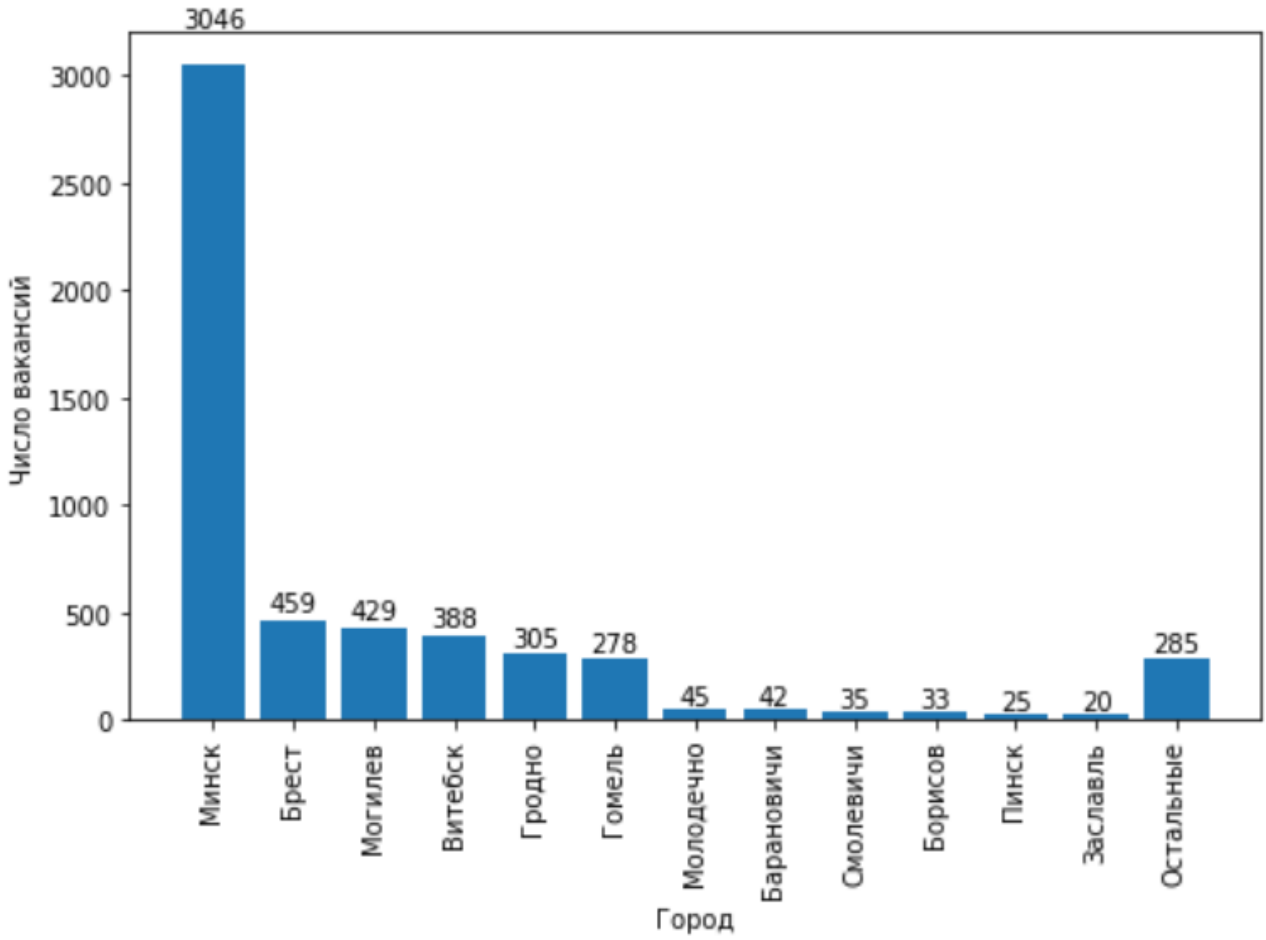

Рисунок 4 – Распределение вакансий по городам

Как и ожидалось, больше половины вакансий предлагается в Минске. Остальные областные центры идут за Минском и предлагают сравнимое (между собой) число вакансий. Остальные города в отдельности предлагают заметно меньшее число вакансий, чем областные центры.

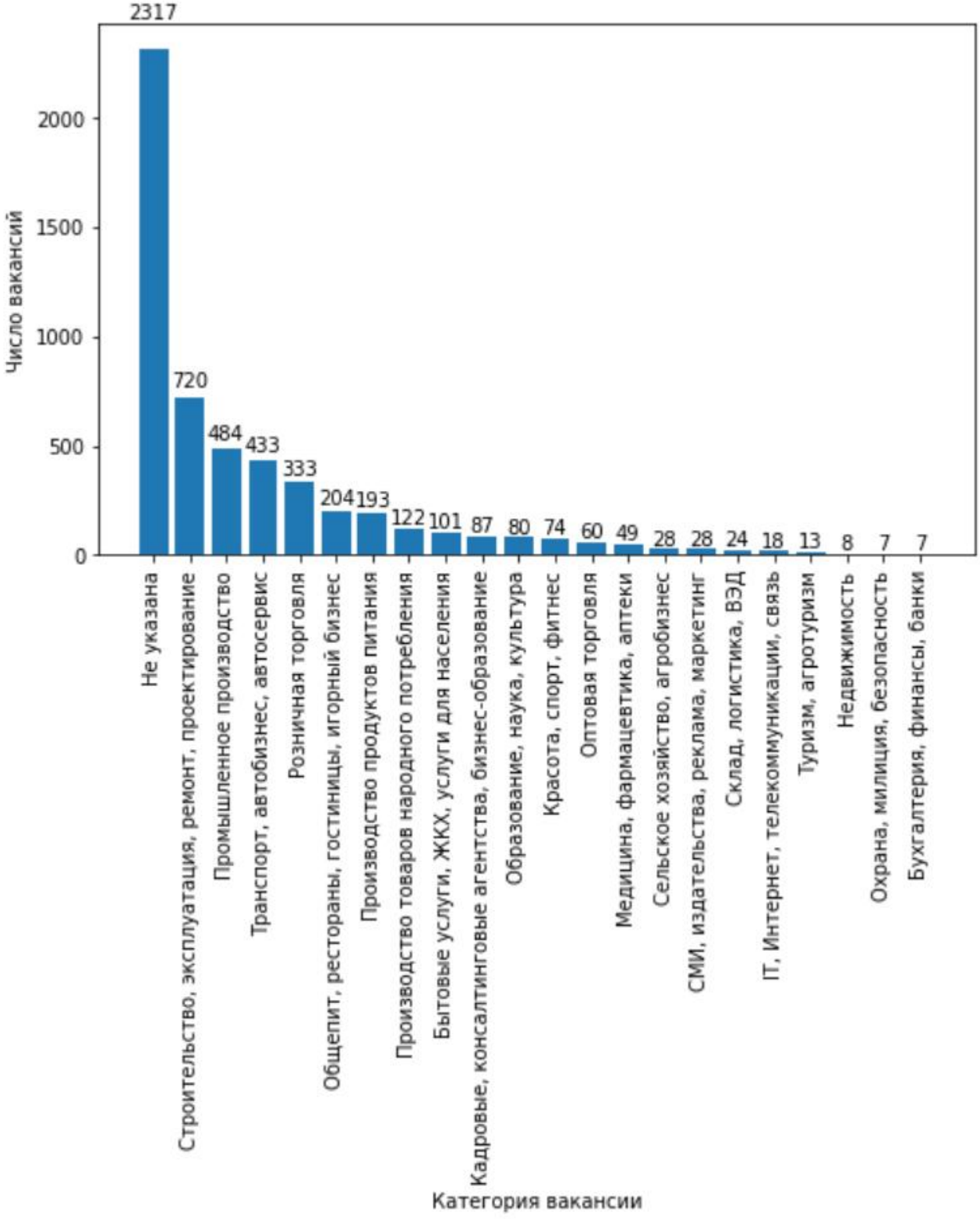

Рисунок 5 – Распределение вакансий по категориям

Для почти половины вакансий не указана их категория. С существенным отрывом лидирует категория "Строительство, эксплуатация, ремонт, проектирование".

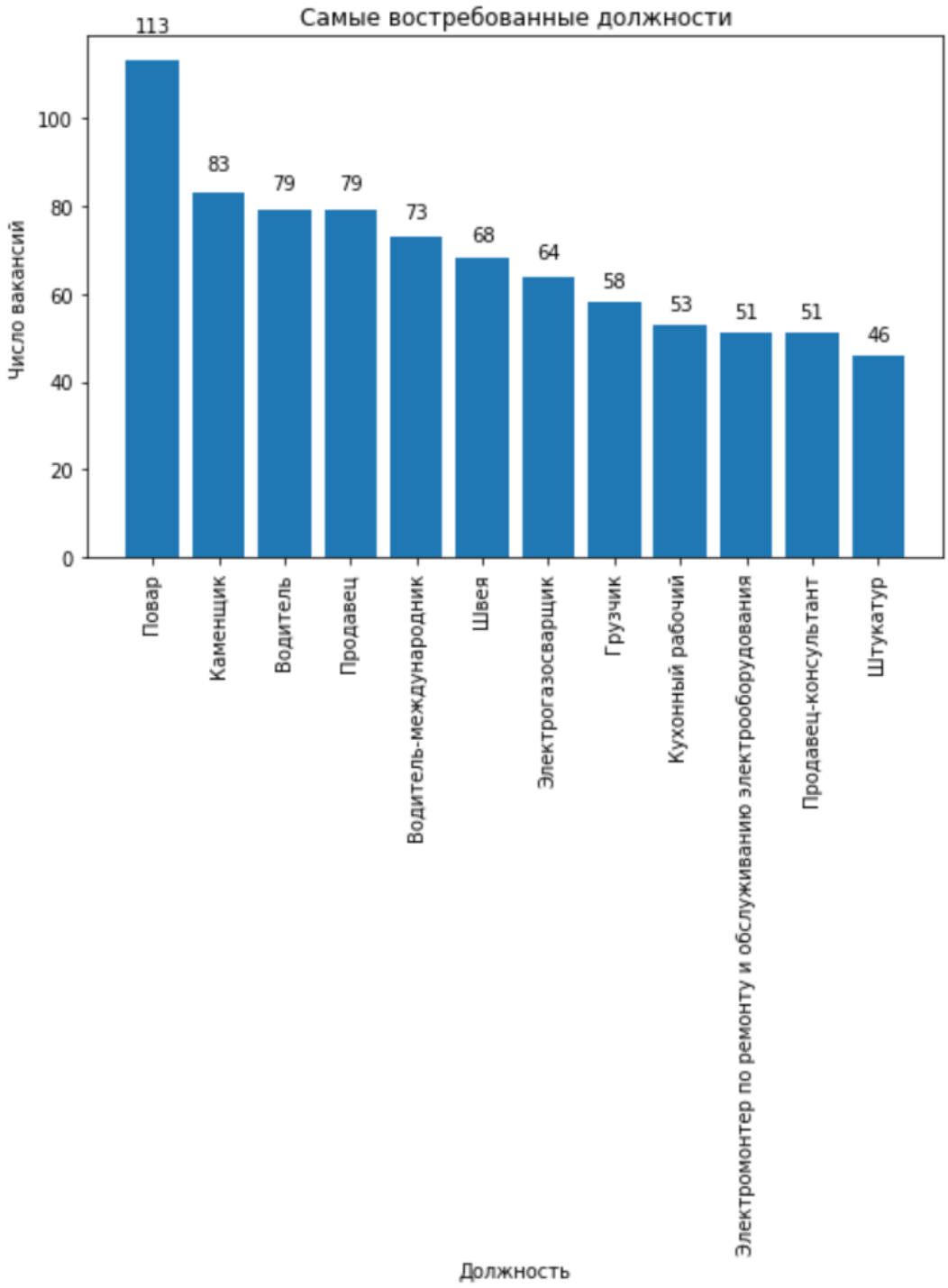

Рисунок 6 – Распределение вакансий по предлагаемым должностям (топ 12 значений)

Если объединить на графике схожие должности "Повар" и "Кухонный рабочий", "Водитель" и "Водитель-международник", "Продавец" и "Продавец-консультант", то в сумме они будут самыми востребованными. Получается, что проще всего найти работу продавцам, водителям и поварам.

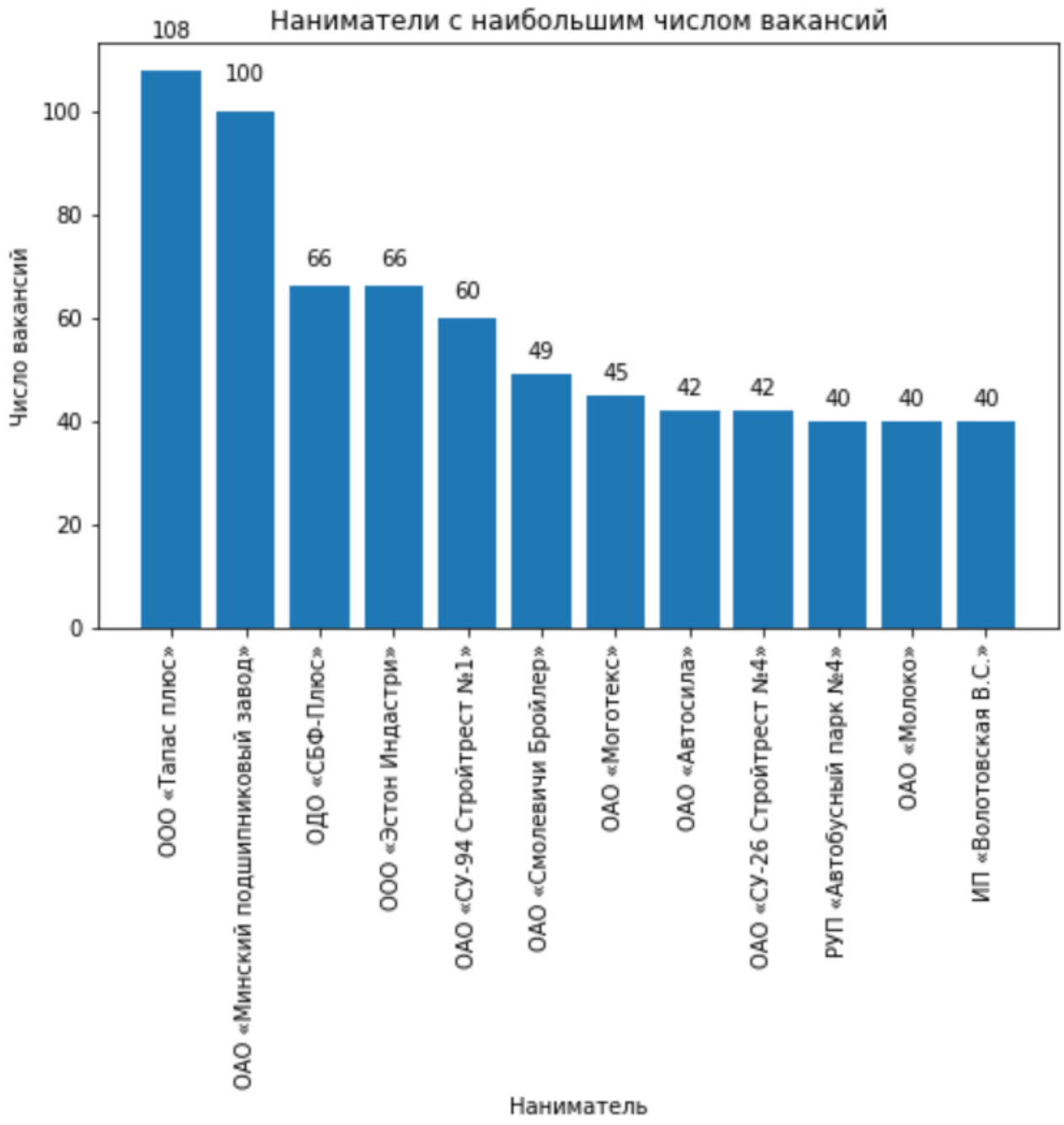

Рисунок 7 – Распределение вакансий по нанимателям (топ 12 значений)

Среди нанимателей с наибольшим числом предлагаемых вакансий больше остальных фигурируют застройщики. Например, ООО "Тапас плюс" и ОДО "СБФ-Плюс".

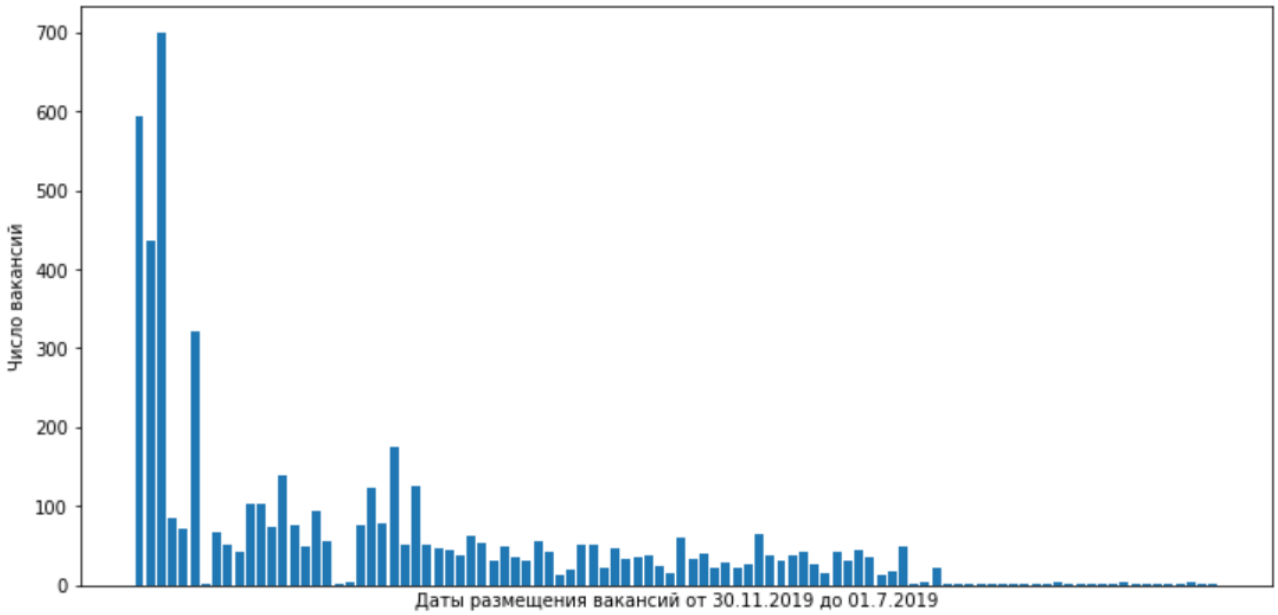

Рисунок 8 – Распределение даты подачи вакансии

Какой-то особой структуры в дате размещения вакансий не выявлено, не считая того, что более "свежих" вакансий больше. Было выявлено, что вакансии не размещаются по воскресеньям.

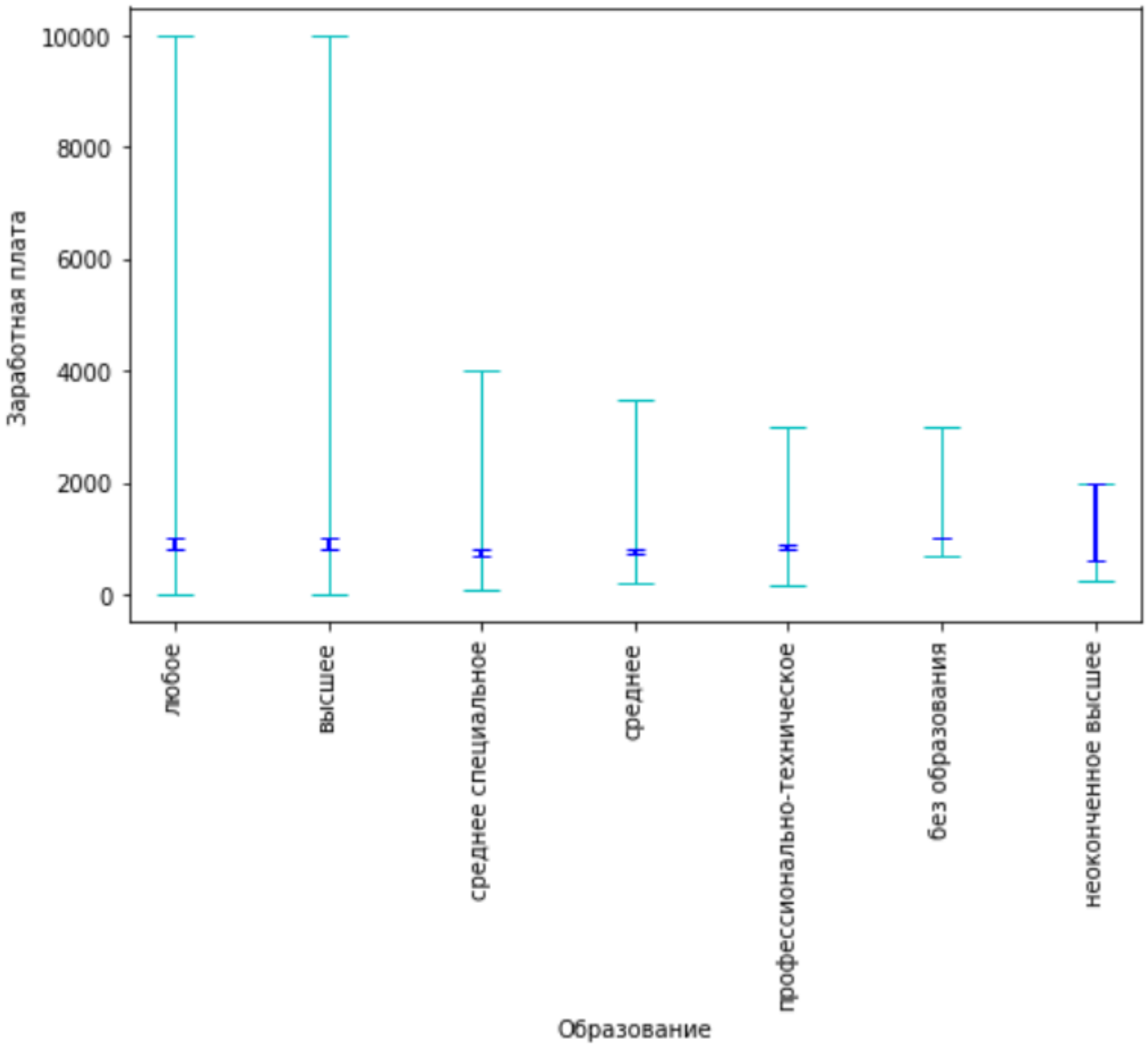

Рисунок 9 – Зависимость абсолютных и медианных границ предлагаемой заработной платы от требуемого уровня образования

Наибольшая заработная плата предлагается на вакансиях, требующих высшее образование, что ожидаемо, и на вакансиях, не требующих конкретного образования, что ожидается меньше. Примером вакансии, предлагающей до 10000 Br, но при этом не требующей образования, является "Водитель". Есть вакансии, которые предлагают всего от 7 Br. Примером такой вакансии является "Преподаватель ландшафтного дизайна". Возможно, в предлагаемой заработной плате опечатка, или плата указана не за месяц, а, например, за одно занятие. Полагаю, что последнее, т.к. ситуация аналогична для других вакансий преподавателей.

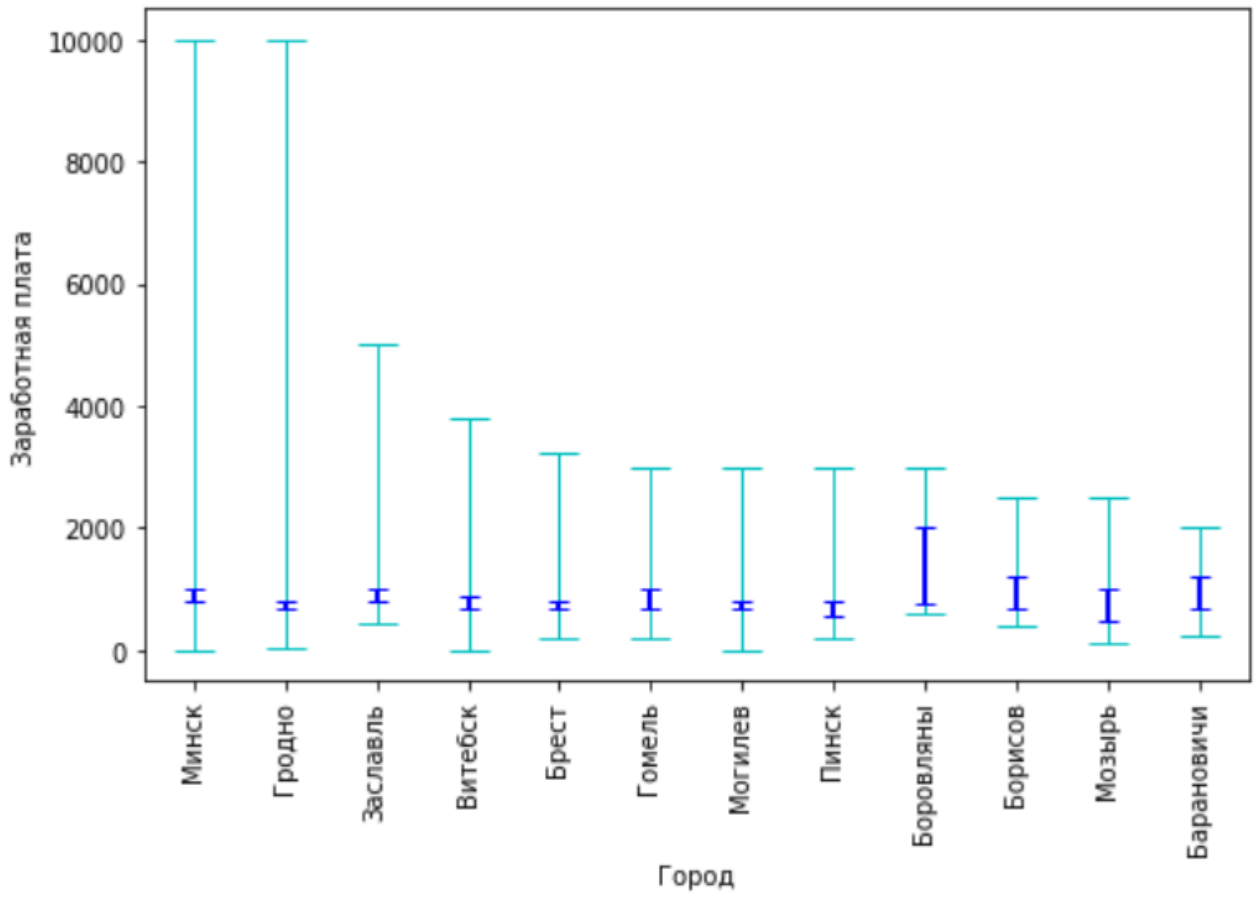

Рисунок 10 – Зависимость абсолютных и медианных границ предлагаемой заработной платы от города (топ 12 значений)

Как и ожидается, в топе городов фигурируют областные центры. Неожиданно туда попал Заславль с вакансией "Вальщик леса", которая предлагает заработную плату в размере от 1000 до 5000 Br. Вакансии "Водитель" и "Проектировщик систем вентиляции и кондиционирования Главный специалист" позволяют Минску и Гродно занять первые строчки. Эти вакансии предлагают заработную плату в размере до 10000 Br.

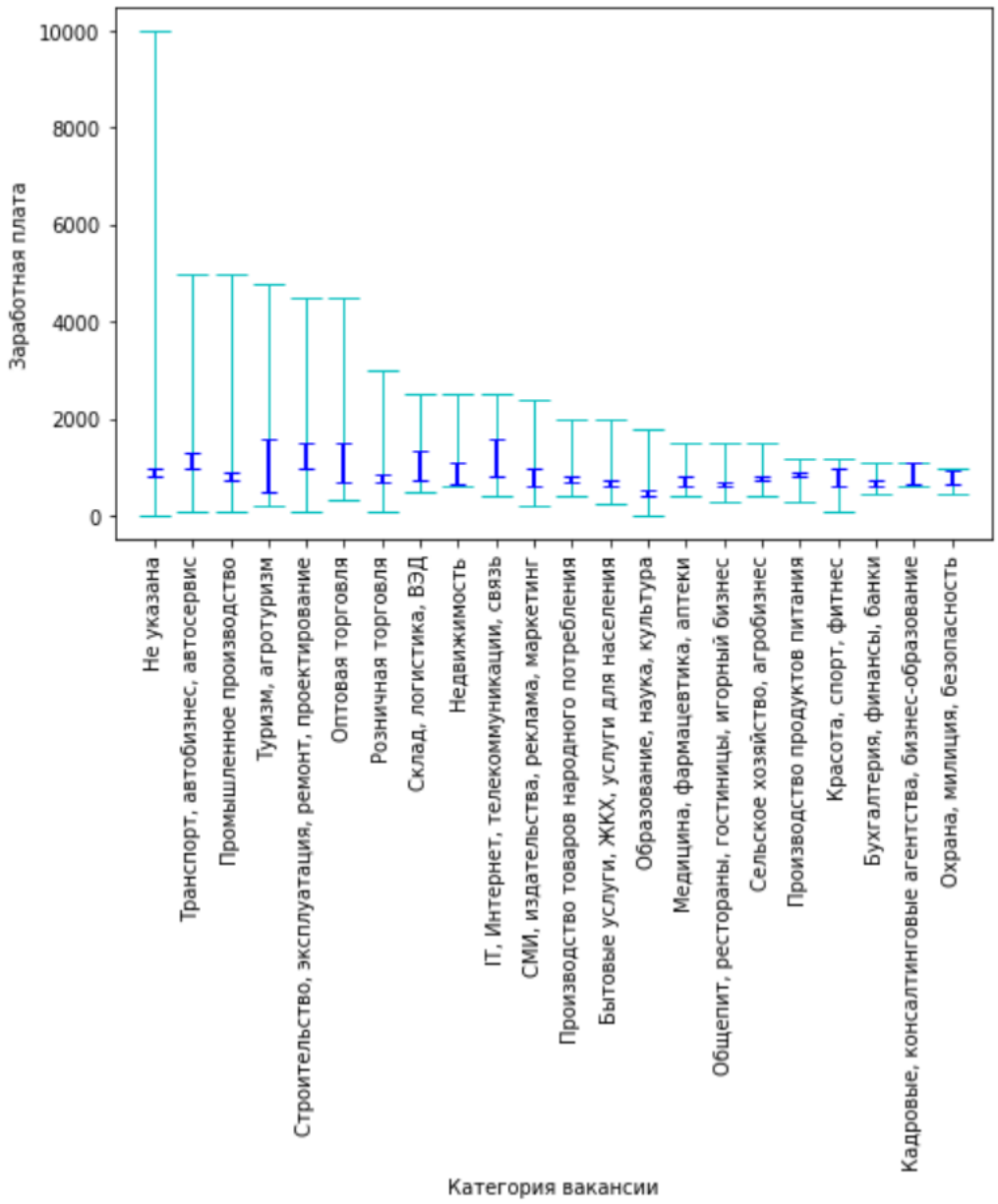

Рисунок 11 – Зависимость абсолютных и медианных границ предлагаемой заработной платы от категории вакансии

Всё те же вакансии "Водитель" и "Проектировщик систем вентиляции и кондиционирования Главный специалист" ставят категорию "Не указана" на первое место. Заметно выше остальных располагаются категории "Транспорт, автобизнес, автосервис", "Промышленное производство", "Туризм, агротуризм", "Строительство, эксплуатация, ремонт, проектирование", "Оптовая торговля". В самом хвосте находятся "Производство продуктов питания", "Красота, спорт, фитнес", "Бухгалтерия, финансы, банки", "Кадровые, консалтинговые агенства, бизнес-образование", "Охрана, милиция, безопасность". Особенно удивительно видеть категорию "Бухгалтерия, финансы, банки" третьей с конца по максимальной предлагаемой заработной плате. От этой категории обычно ожидаются высокие доходы.

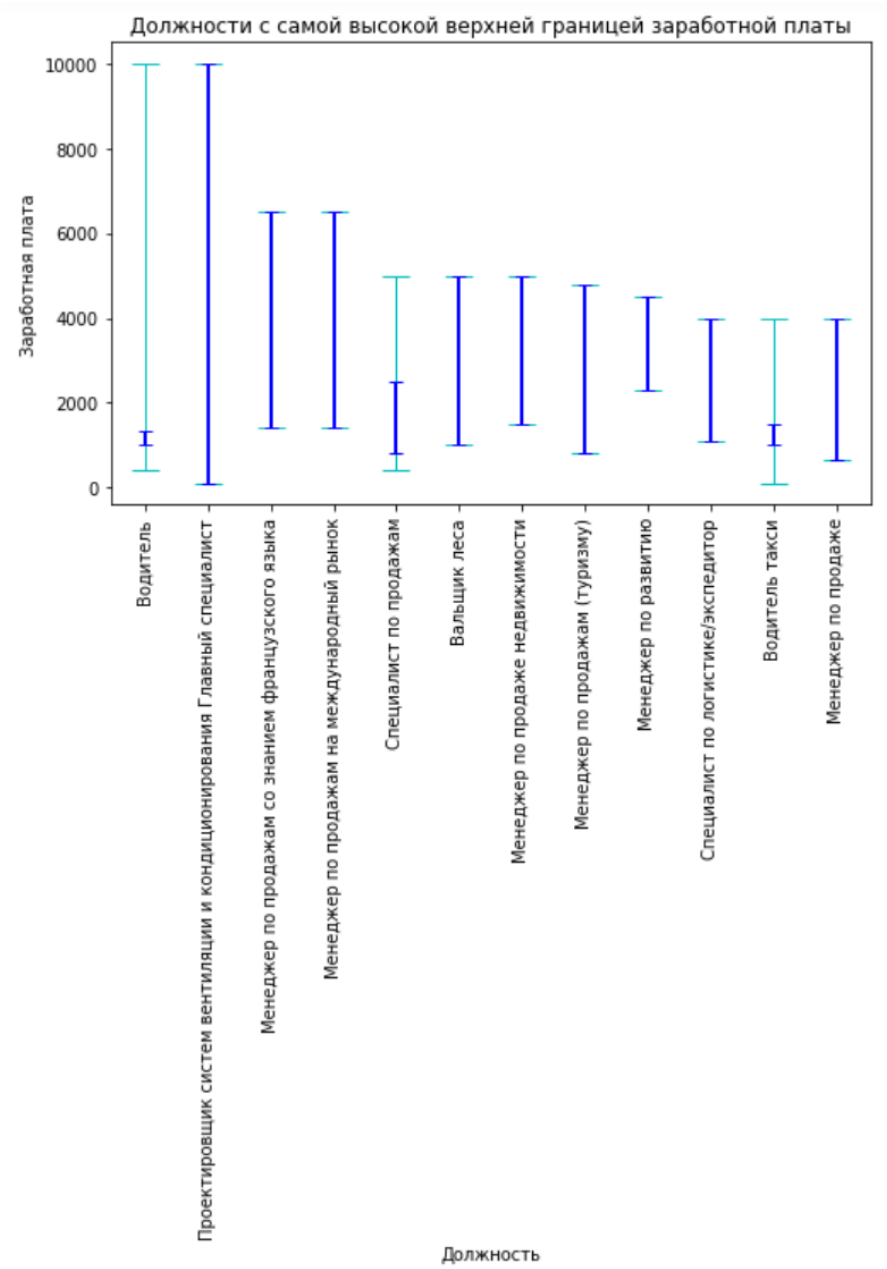

Рисунок 12 – Зависимость абсолютных и медианных границ предлагаемой заработной платы от предлагаемой должности (топ 12)

Интересно видеть, что должности с самой высокой возможной заработной платой также имеют очень низкую нижнюю границу. Заработная плата вакансии "Проектировщик систем вентиляции и кондиционирования Главный специалист" может варьироваться от 100 до 10000 Br, т.е. верхняя граница отличается от нижней в 100 раз. Похожая ситуация и у вакансии "Водитель". Вакансиями с высокой не только верхней, но и нижней границей заработной платы являются вакансии менеджеров, например "Менеджер по продажам со знанием французского языка". Стоит отметить, что среди должностей с самой высокой верхней границей заработной платы много менеджеров. Также заметим, что вакансии водителей имеют несравнимо меньшие медианные границы заработной платы.

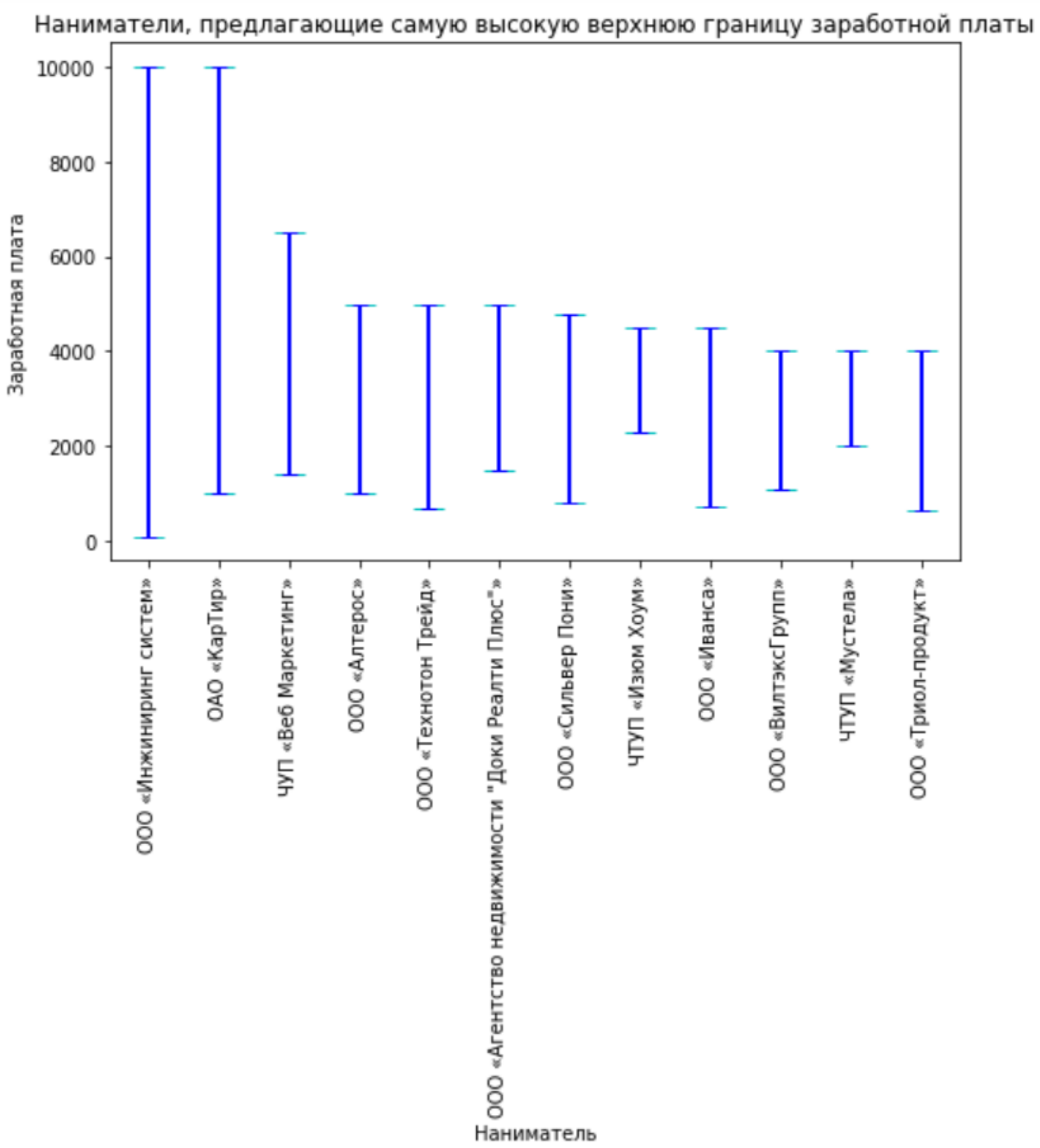

Рисунок 13 – Зависимость абсолютных и медианных границ предлагаемой заработной платы от нанимателя (топ 12)

ООО "Инжиниринг систем" ищет того самого "Проектировщик систем вентиляции и кондиционирования Главный специалист" с заработной платой до 10000 Br, а ОАО "КарТир" – того самого водителя с аналогичной верхней границей заработной платы. ЧУП "Веб Маркетинг" ищет менеджеров: "Менеджер по продажам со знанием французского языка", "Менеджер по продажам на международный рынок".

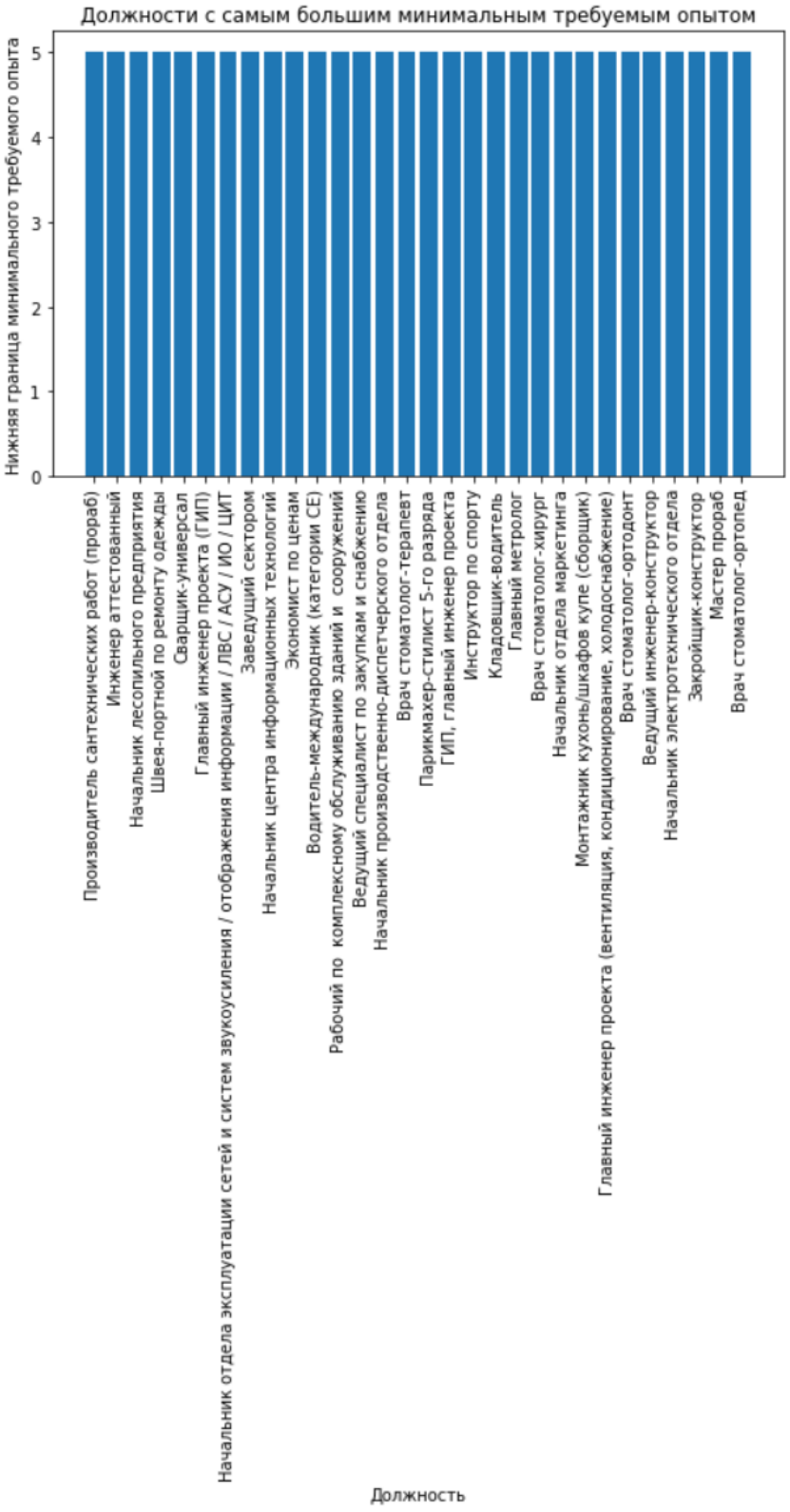

Рисунок 14 – Зависимость минимального требуемого опыта от предлагаемой должности (топ 30)

Наибольший минимальный требуемый опыт составляет 5 лет. Среди вакансий, требующих такой опыт, много начальников, ведущих специалистов и, что интересно, врачей-стоматологов. Медианная минимальная заработная плата для вакансий, требующих такой опыт, составляет 1000 Br.

Посчитанные абсолютные границы предлагаемой заработной платы: от 7 до 10000 Br.

Посчитанные медианные границы предлагаемой заработной платы: от 800 до 1000 Br.

Медианная зарплата по Беларуси за май 2019-го года составила 802,7 рубля, что соответствует посчитанным медианным границам. Полагаю, людям значительно чаще платят ближе к минимальной границе предлагаемой заработной платы. Высокие верхние границы заработной платы могут использоваться для привлечения внимание.

Выводы

Результат выполнения лабораторной работы: при помощи языка программирования Python и фреймворка Scrapy выгружены данные о вакансиях с сайта https://rdw.by/vakansii, при помощи стандартной библиотеки языка программирования Python и библиотеки matplotlib выполнены обработка и анализ выгруженных данных, построены графики. Поставленная задача выполнена полностью.

В рамках анализа данных о вакансиях было изучено распределение вакансий по требуемому образованию, городам, категориям, предлагаемым должностям, нанимателям, дате размещения, исследован требуемый для вакансии опыт, а также размер предлагаемой заработной платы. Посчитанный медианный уровень предлагаемой в вакансиях заработной платы соответствует официальной статистике, предоставляемой Национальным статистическим комитетом Беларуси. При этом сделан вывод, что людям значительно чаще платят ближе к минимальной границе предлагаемой заработной платы.

Приобретенные в рамках выполнения лабораторной работы навыки:

1. Научился выполнять выгрузку данных с сайтов с использованием языка программирования Python и фреймворка Scrapy.
2. Развил навыки работы с библиотекой matplotlib, используемой для построения графиков.
3. 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), datetime (работа с датой и временем).