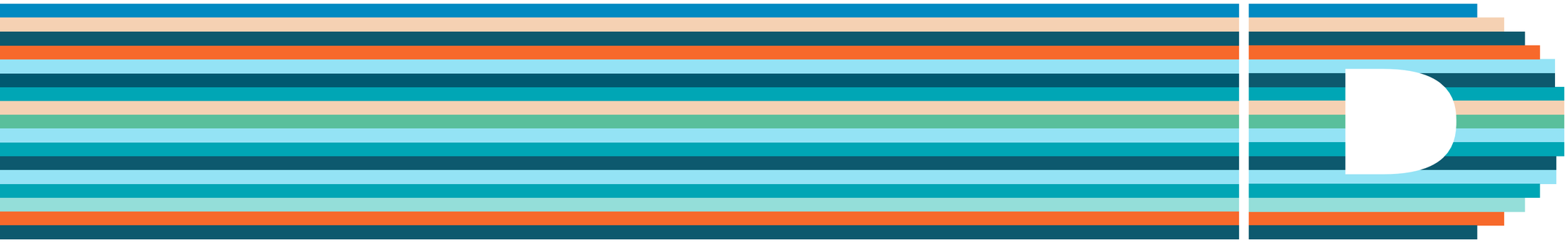
Итоговый проект по программе «Data-аналитик: старт карьеры»

Поток № DA-104— сентябрь 2022 г.



Национальный исследовательский Томский государственный университет

data-diving

академия аналитики данных при Томском государственном университете

Описание проекта

Название проекта: **Анализ рынка труда в сфере IT и рынка онлайн-курсов в сфере IT.**

Бизнес-цель заказчика: Запустить образовательные продукты для IT-специалистов: онлайн-курсы повышения квалификации.

Объект исследования: 1. Рынок образовательных услуг в сфере IT.
2. Рынок труда в сфере IT.

Предмет исследования: 1. Количество программ в сфере онлайн-курсов, их тематическая направленность, продолжительность обучения, их стоимость и средняя оценка потребителей. 2. Требования работодателей к соискателям на позиции IT-специалистов разных направлений. Информация о средней заработной плате для специалистов в данной области, количество вакансий.

Описание проекта

Цель исследования:

1. Выявить наиболее востребованные и актуальные Образовательные Программы на рынке труда в сфере IT.
2. Актуализировать соотношение спроса и предложения на рынке труда в сфере IT.

Описание проекта

Требования к результату анализа:

1. Наглядное описание ситуации на рынке труда и ситуации в дополнительном образовании (онлайн-курсы);
2. Результат бенчмаркинга продуктов конкурентов;
3. Представление выводов и рекомендаций:
   - какие образовательные продукты нужны рынку,
   - какую цену (или диапазон цен) можно поставить, чтобы не продешевить, но и люди могли покупать.

* Риски и условия реализации проекта (факультативно):

1. Плохое качество или недостаточное количество данных.
2. В данных отсутствует необходимая информация.

Описание проекта

Источники данных, типы данных:

Сайты с полуструктурированными данными:

- Tutortop.ru – один из крупных российских агрегаторов онлайн-курсов (тип html);
- HH.ru – крупнейший российский сервис, который помогает найти работу и подобрать персонал (тип json).

Способ(ы) получения данных:

Сбор открытых данные данных с помощью:

- Парсинг сайта Tutortop.ru, разбор html с помощью библиотеки BeautifulSoup;
- Получение данных с использованием API сайта HH.ru.

Описание проекта

Этапы исследования

Планирование дизайна исследования:

Составлен план исследования: определены объекты исследования. После Декомпозиции объектов исследования были получены предметы для исследования их свойств и характеристик. Поставлена цель анализа данных и определены требования к результату анализа.

*Сбор данных:

Через API сайта HH.ru, возможна загрузка 2000 вакансий в одном поисковом запросе: решено загружать вакансии по работодателю, но API отдает 5000 работодателей в запросе, загружен список работодателей с открытыми вакансиями по городам. Вакансии работодателя загружались по временным интервалам, чтобы не было превышения 2000 вакансий в запросе.

Описание проекта

***Сбор данных:** Ежедневно с сайта HH.ru выгружались данные по количеству вакансий и резюме в разрезе специализаций и навыкам.

Данные с сайта Tutortop.ru были получены обычным парсингом.

Обработка данных:

На этапе разведочного анализа данных была выполнена очистка данных от выбросов, заполнены пропуски в данных и выделены ключевые навыки вакансий и специализации в отдельные поля.

Статистическое исследование данных:

- Описательные статистики применялись для количественного описания данных с помощью основных статистических показателей.
- Корреляционный анализ применялся для измерения взаимосвязи между признаками.

Описание проекта

Интерпретация данных:

Использовались методы неграфического (табличного) и графического анализа данных.

Оформление результатов анализа:

Результаты анализа данных отображались с помощью гистограмм, линейных графиков, диаграмм рассеяния, тепловых карт для матрицы корреляции признаков, боксплотов (ящиков с «усами») и круговых диаграмм.

Результаты анализа

Гипотезы исследования:

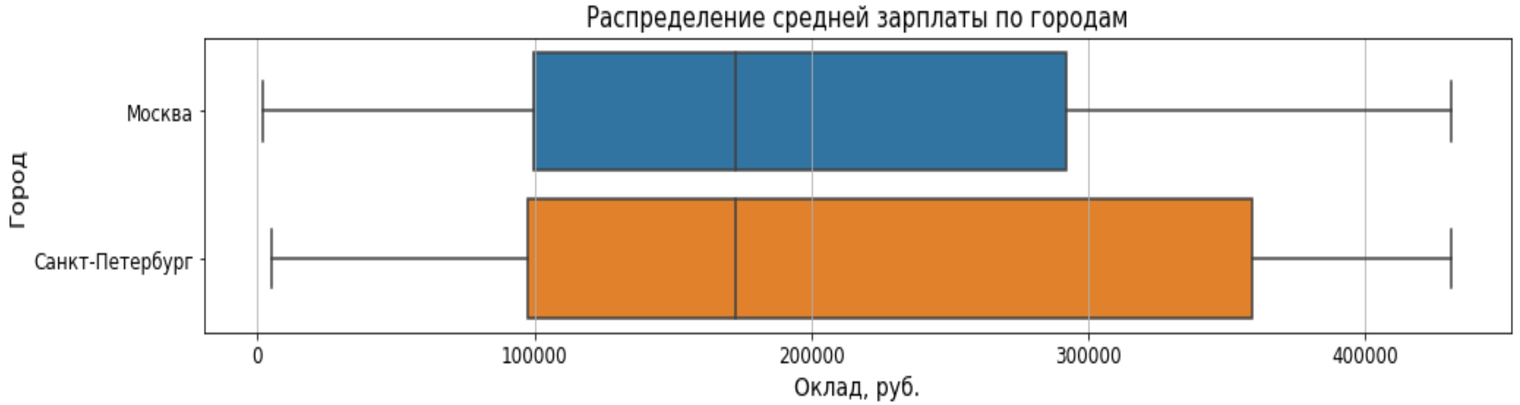
1. Средняя зарплата по вакансиям "Аналитик данных" не отличается в Москве и Санкт-Петербурге.
2. Средняя стоимость курсов в ИТ категориях не отличается.

Метод(ы) проверки гипотез:

1. Проверка вида распределения зарплаты и стоимости курсов на нормальность.
2. В зависимости от результата П.1 выбираем критерий для сравнения групп.

Результаты анализа

Результаты проверки гипотезы (графическое представление)



Проверка нормальности распределения по критерию Шапиро-Уилка:

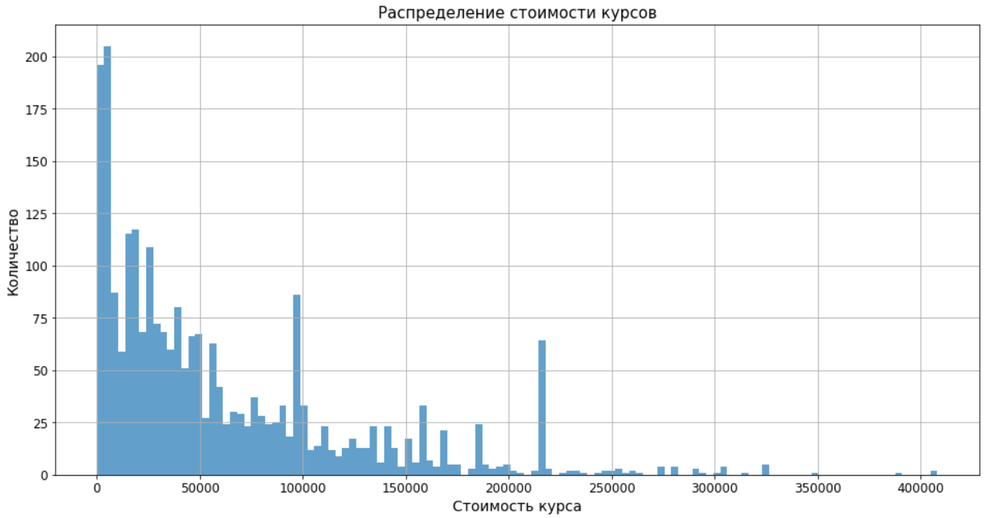
p-value = 0.0025, что меньше уровня значимости $\alpha = 0.05$

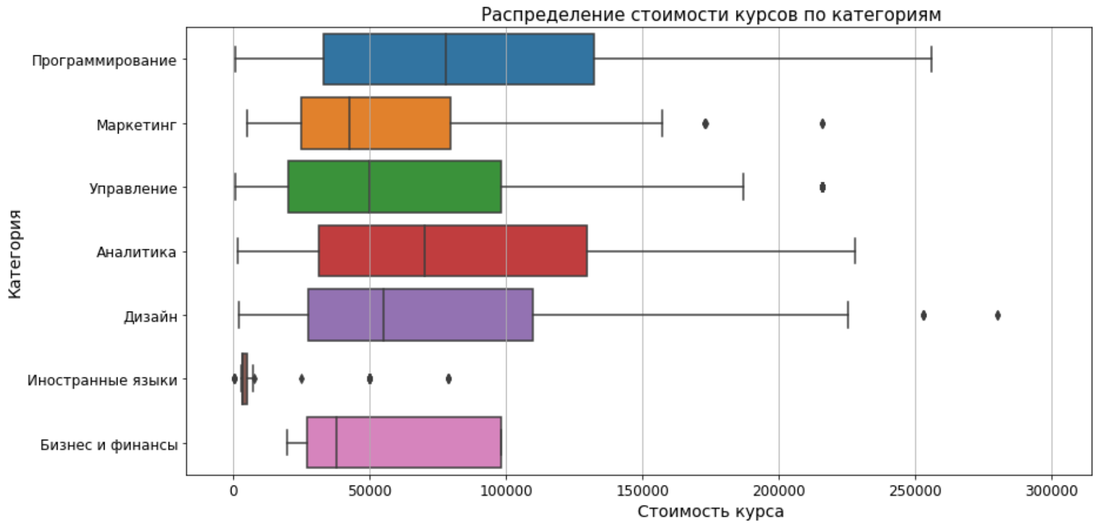
Отвергаем нулевую гипотезу о нормальности распределения и для сравнения групп используем критерий Манна — Уитни:

p-value = 0.3566, что больше уровня значимости $\alpha = 0.05$

Принимаем нулевую гипотезу о равенстве средней зарплаты в Москве и Санкт-Петербурге.

Результаты анализа





Проверка нормальности распределения стоимости курсов по критерию Шапиро-Уилка:

p-value = 0.0000, что меньше уровня значимости $\alpha = 0.05$

Отвергаем нулевую гипотезу о нормальности распределения и для сравнения групп используем критерий Крускала-Уоллиса (больше 2-х групп, данные ненормальные) :

p-value = 0.0148, что меньше уровня значимости $\alpha = 0.05$

Отвергаем нулевую гипотезу об одинаковой стоимости курсов в разных категориях.

Сравнения стоимости курсов в категориях «Программирование» и «Аналитика»:

p-value = 0.3756 - Принимаем нулевую гипотезу об одинаковой стоимости курсов.

Интерпретация данных

Полученные результаты:

Анализ предложений по вакансиям сайта HH.ru (ссылка на дашборд) показал, что ИТ-специалисты находится на 7-м месте в рейтинге востребованности профессий.

Спрос на предложения работодателей оценивался по статусам в резюме «Активно ищет работу» и «Рассматривает предложения», статусы резюме «Без статуса поиска» и «Не ищет работу» – в анализ не включались.

На основе вакансий были составлены рейтинги востребованности ИТ-специалистов по группам (Тип поиска):

- Наименование вакансии
- Специализация
- Ключевой навык

Из каждой группы были отобраны по 15 не пересекающихся позиций рейтинга, если какая-то позиция входила в разные группы – то из группы с меньшим рейтингом она исключалась.

Интерпретация данных

Полученные результаты:

На основе данных по онлайн-курсам сайта Tutortop.ru отмечены крупные Школы – авторы курсов по совокупности признаков: количество курсов, количество отзывов, рейтинг пользователей.

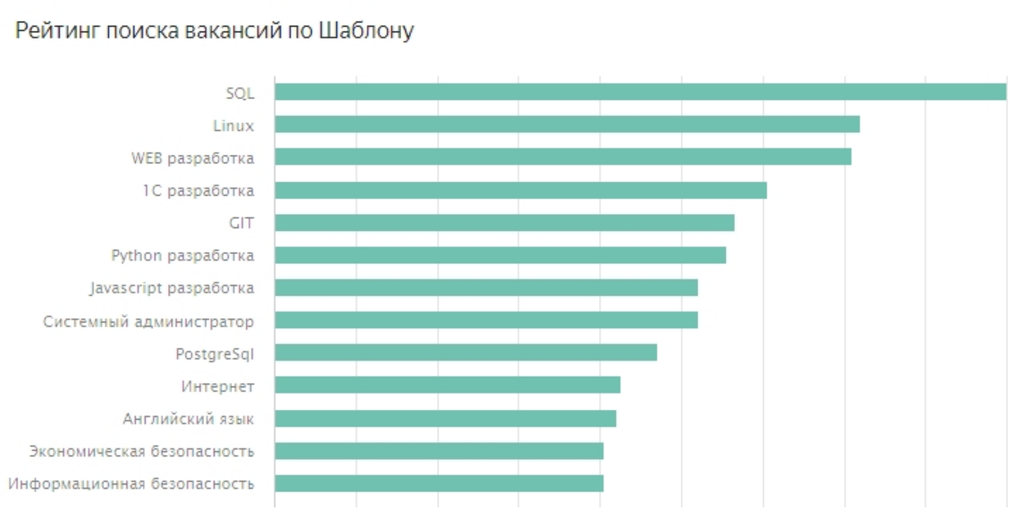
На основе групп востребованности ИТ-специалистов осуществлен поиск по списку курсов, исключая разделы "Детям", "Образ жизни" и "Создание контента". При подборе списка курсов для вакансий, специализаций и навыков использовалась библиотека NLTK (обработка естественного языка): выделение ключевых слов, их лемматизация и стемминг для вычисления коэффициента похожести текста. Порог похожести текста был принят 95%.

Для каждого шаблона поиска посчитано количество найденных курсов, их средняя стоимость и продолжительность обучения (ссылка на дашборд).

Интерпретация данных

Полученные результаты:

Для специализаций "Системы управления предприятием (ERP)" и "Поддержка, Helpdesk" (1 и 3 место рейтинга) не найдено ни одного курса. Анализ ключевых навыков этих специализаций показал, что все позиции из первой десятки навыков присутствуют в Рейтинге поиска вакансий по Шаблону в соответствующих пропорциях, но в другой последовательности в рейтинге.



Интерпретация данных

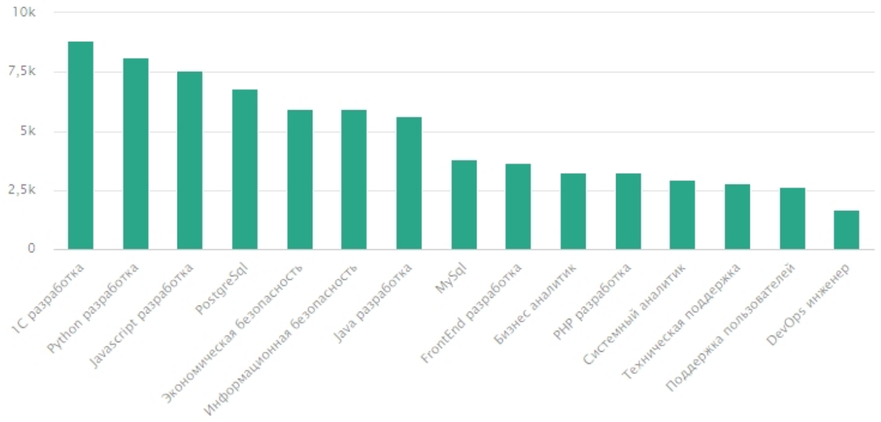
Рекомендации для Заказчика:

На основе дашборда для лидеров рейтинга сформировать курсы, указав стоимость выше средней на 10-15%, например «Программирование 1С»

Тип поиска: Вакансия

Крупная школа: —

ТОП-15 позиций сайта hh.ru



Шаблон: 1С разработка

Позиция	Шаблон поиска курсов	Кол-во курсов	Длит, мес	Средняя цена
1	1С разработка	5	10,80	107 600
2	Python разработка	59	5,70	81 300
3	Javascript разработка	58	6,00	70 400
4	PostgreSql	1	1,50	35 000
5	Информационная безопасность	26	3,90	69 500
6	Экономическая безопасность	2	0,00	131 900
7	Java разработка	45	7,40	96 300
8	MySql	1	1,50	16 500
9	FrontEnd разработка	28	10,20	95 200
10	Бизнес аналитик	55	2,60	85 200
11	PHP разработка	24	6,80	61 500
12	Системный аналитик	22	2,80	101 500
13	Техническая поддержка	0	*null*	*null*
14	Поддержка пользователей	0	*null*	*null*
15	DevOps инженер	5	10,60	101 200

Наименование курса	Длит, мес	Цена	Школа курса
Профессия 1С-разработчик	6,00	61 596	Skillbox
Профессия Аналитик 1С	15,00	87 504	Skillbox
1С-программист	11,00	90 000	Нетология
Факультет 1С-разработки	10,00	112 068	GeekBrains
Разработчик: специализация 1С-разработка	12,00	186 876	GeekBrains

Интерпретация данных

Перспективные направления для дальнейшего анализа:

Сформировать курсы на основе отношения долей вакансий/курсов, т.е. где на ранке мало курсов и есть спрос на специалистов с соответствующими навыками:

Тип поиска — Фильтры: Доля вакансий (операция >=), % 5 Доля курсов (операция >=), % 0.1

Тип поиска	Позиция	Шаблон поиска	Доля вакансий, %	Доля курсов, %	Отношение долей: вакансий/курсов
Вакансия	1	1С разработка	12,10	1,50	80,30
Вакансия	4	PostgreSql	9,40	0,30	77,50
Ключевой навык	5	Работа в команде	6,20	0,40	30,40
Вакансия	6	Экономическая безопасность	8,10	0,60	22,40
Вакансия	8	MySql	5,20	0,30	21,60
Ключевой навык	1	SQL	18,00	9,70	18,50
Специализация	4	Системный администратор	10,40	1,90	14,00
Ключевой навык	2	Linux	14,40	6,90	10,50
Ключевой навык	3	GIT	11,30	3,60	10,30
Специализация	2	WEB разработка	14,20	17,00	4,20
Вакансия	2	Python разработка	11,10	17,80	3,10
Вакансия	5	Информационная безопасность	8,10	7,90	2,10
Вакансия	3	Javascript разработка	10,40	17,50	2,00
Специализация	5	Интернет	8,50	9,80	1,70
Вакансия	7	Java разработка	7,70	13,60	0,80
Ключевой навык	4	Английский язык	8,40	27,90	0,80
Вакансия	9	FrontEnd разработка	5,00	8,50	0,70
Специализация	8	Аналитик	5,70	15,60	0,50

Интерпретация данных

Перспективные направления для дальнейшего анализа:

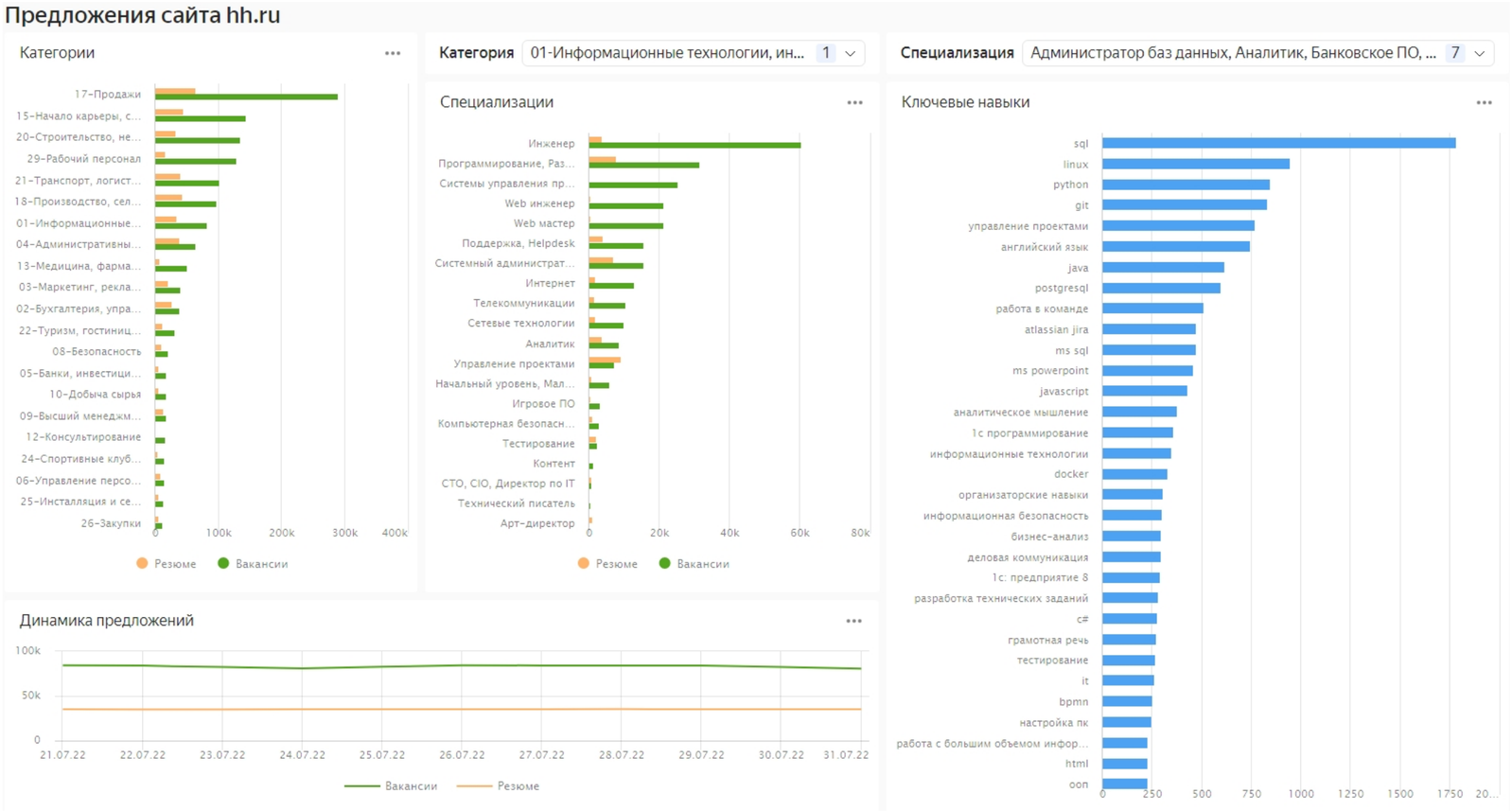
Исследование динамики размещения вакансий и резюме на сайте hh.ru для ИТ-специализаций (на текущий момент недостаточно данных, всего 5 недель):

Проверка временных рядов на стационарность с помощью теста Дики-Фуллера:

- для вакансий: p-value = 0.2168
- для резюме: p-value = 0.5063

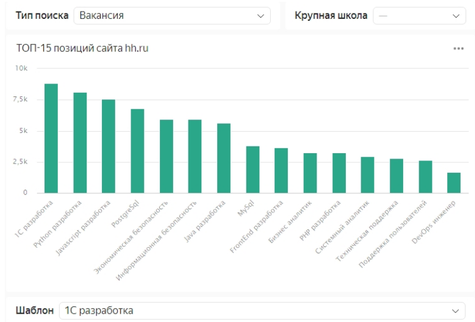
В обоих случаях принимаем нулевую гипотезу о не стационарности временного ряда.
Анализ графиков выявил для вакансий и резюме тренд увеличения количества предложений с течением времени.
Динамика публикаций имеет сезонный характер с периодом в неделю, но разной точкой отсчета для вакансий и резюме.
Подробности тут.

Дашборд 1



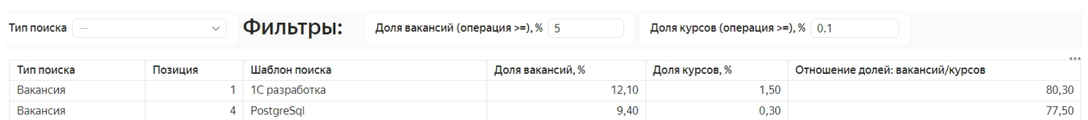
Информация о распределении вакансий и резюме сайта hh.ru по категориям специализаций: ТОП 20 специализаций в выбранной категории, ТОП-30 ключевых навыков для выбранных специализаций. Присутствует динамика количества вакансий и резюме по датам.
Фильтры:
- выбор категорий (одной, нескольких или всех)
- выбор специализаций (одной, нескольких или всех)

Дашборд 2



Позиция	Шаблон поиска курсов	Кол-во курсов	Длит, мес	Средняя цена
1	1С разработка	5	10,80	107 600
2	Python разработка	59	5,70	81 300
3	Javascript разработка	58	6,00	70 400
4	PostgreSql	1	1,50	35 000
5	Информационная безопасность	26	3,90	69 500
6	Экономическая безопасность	2	0,00	131 900
7	Java разработка	45	7,40	96 300
8	MySql	1	1,50	16 500
9	FrontEnd разработка	28	10,20	95 200
10	Бизнес аналитик	55	2,60	85 200
11	PHP разработка	24	6,80	61 500
12	Системный аналитик	22	2,80	101 500
13	Техническая поддержка	0	*null*	*null*
14	Поддержка пользователей	0	*null*	*null*
15	DevOps инженер	5	10,60	101 200

Наименование курса	Длит, мес	Цена	Школа курса
Профессия 1С-разработчик	6,00	61 596	Skillbox
Профессия Аналитик 1С	15,00	87 504	Skillbox
1С-программист	11,00	90 000	Нетология
Факультет 1С-разработки	10,00	112 068	GeekBrains
Разработчик: специализация 1С-разработка	12,00	186 876	GeekBrains



Тип поиска	Позиция	Шаблон поиска	Доля вакансий, %	Доля курсов, %	Отношение долей: вакансий/курсов
Вакансия	1	1С разработка	12,10	1,50	80,30
Вакансия	4	PostgreSql	9,40	0,30	77,50
Ключевой навык	5	Работа в команде	6,20	0,40	30,40
Вакансия	6	Экономическая безопасность	8,10	0,60	22,40
Вакансия	8	MySql	5,20	0,30	21,60
Ключевой навык	1	SQL	18,00	9,70	18,50
Специализация	4	Системный администратор	10,40	1,90	14,00
Ключевой навык	2	Linux	14,40	6,90	10,50
Ключевой навык	3	GIT	11,30	3,60	10,30
Специализация	2	WEB разработка	14,20	17,00	4,20
Вакансия	2	Python разработка	11,10	17,80	3,10
Вакансия	5	Информационная безопасность	8,10	7,90	2,10
Вакансия	3	Javascript разработка	10,40	17,50	2,00
Специализация	5	Интернет	8,50	9,80	1,70
Вакансия	7	Java разработка	7,70	13,60	0,80
Ключевой навык	4	Английский язык	8,40	27,90	0,80
Вакансия	9	FrontEnd разработка	5,00	8,50	0,70
Специализация	8	Аналитик	5,70	15,60	0,50

- Информация о курсах tutortop.ru – Просмотр ТОП-15 позиций. Фильтры:
  1. «Типа поиска» выбирает статистику по Вакансиям, Специализациям и Ключевым навыкам
  2. «Крупная школа» - фильтрует статистику «Все школы», «Крупные», «Кроме крупных»
  3. «Шаблон» - фильтрует курсы по шаблону поиска
- Специализации без курсов – Просмотр информации по ключевым навыкам специализаций для которых не было найдено ни одного курса.
- Рейтинг отношений Вакансий и Курсов – Просмотр ТОП-15 позиций. Фильтры:
  1. «Типа поиска» выбирает статистику по Вакансиям, Специализациям и Ключевым навыкам
  2. «Доля вакансий» - фильтрует статистику, где доля вакансий выше определенного %
  3. «Доля курсов» - фильтрует статистику, где доля курсов выше определенного %

* Дополнительно (ссылки на материалы)

Блокнот: EDA – Разведочный анализ данных сайта hh.ru на данных заказчика
https://colab.research.google.com/drive/1ys7PTaJhb0FiXnb3lNNY7qdEr6h10_zu?usp=sharing

Блокнот: Проверка статистической гипотезы
https://colab.research.google.com/drive/1At9Or-ndM2KAPFKVJbqxRHt6N5x7B09f?usp=sharing

Блокнот: EDA - разведочный анализ данных на собственных данных с hh.ru - Исследование ИТ направлений
https://colab.research.google.com/github/saspav/DA-104/blob/main/Павлова_СВ_EDA_hh.ipynb

Блокнот: Динамика размещения вакансий и резюме на сайте hh.ru для ИТ-специализаций
https://colab.research.google.com/github/saspav/DA-104/blob/main/hh_vacancy_resume_stats.ipynb

Блокнот: EDA - разведочный анализ данных сайта tutortop.ru
https://colab.research.google.com/drive/1rTdOvjd6VKtJgEUmVXJzGwIbp6ds8FFi?usp=sharing

Дашборд № 1: https://datalens.yandex/2zr6j2bj2scit

Дашборд № 2: https://datalens.yandex/97zam19fo8is0

Дашборд № 3 (на собственных данных): https://datalens.yandex/qxsrhrgr15z6h

Ссылка на репозитарий: https://github.com/saspav/DA-104.git

Спасибо
за внимание!