# Hackathon Masterskaya Practicum

---

Техническое задание на анализ и визуализацию целей обучения

студентов “Яндекс Практикум” БИЗНЕС-ТРЕБОВАНИЯ

Информация о компании

1. Отрасль и направления деятельности: EdTech, сервис-онлайн образования
2. О проекте: Создание и оформление отчёта целей обучения студентов Яндекс Практикума для презентации топ-менеджменту Яндекс Практикума.
3. Цели отчёта:

● определить нормальные и найти аномальные показатели

● определить коррелирующие параметры, построить портреты студентов, сравнить их, чтобы выделить значимые закономерности

● сегментировать студентов (по 2м и более показателям), выявить особенности сегментов

● сформулировать на основе данных гипотезы по улучшению выстраивания помощи студентам в достижении их целей,

● оформить выводы и гипотезы аналитиков с помощью инструментов фигмы для презентации руководству Яндекс Практикума.

## Описание данных

data_goals_answers

- question_title — текст вопроса
- question_type — тип вопроса
- user_id — уникальный id пользователя
- user_answer — ответ пользователя на вопрос
- answer_date — время ответа
- answer_id — id ответа
- cohort, current_cohort — начальная и текущая когорта студента
- course_name, topic_name, lesson_name — курс, тема и урок, на котором студент отвечает на вопрос У нас значения должны быть Трудоустройство-Трудоустройство-Цель обучения, т.к. мы изучаем именно это
- original_segment, current_segment — b2c/b2b/b2g — из какого сегмента был/стал студент — сам является клиентом, его обучение оплачивается бизнесом или государством
- profession_name — код профессии
- statement_content — формулировка вопроса об уверенности в знаниях (в этой таблице нету)
- slide_position — страница опроса (не нужно для анализа)

hackathon_metrics

- profession_name — код профессии
- user_id — уникальный id пользователя
- lp_avg_user — средний learning performance Первые, более высокие значения в таблице с фри-трека, последние с курса, наиболее актуально находящееся в таблице ниже
- question_title — текст вопроса
- user_answer — ответ пользователя на вопрос
- statement_content — формулировка вопроса об уверенности в знаниях
- value — ответ на вопрос об уверенности в знаниях для расчёта learning experience индекса

**Результаты исследования**

В исследование принимали участие только студенты курсов, дающих возможность получить специальность. Курсы математика для da/ds, datavis-and-bi-tools являются дополнительными к основным программам, количество респондентов для таких курсов в наших данных очень незначительно, поэтому некоторые исследования проводились не в рамках курсов, а в рамках специальности, которую студент получает.

Стоит отметить, что в таблице data_goals_answers присутствуют ответы только 14 человек со специальности инженер данных, этот факт делает невозможным какие-либо обобщающие выводы для всего курса ввиду недостаточного размера выборки

**1 Особенности данных (включая аномальные и редкие значения)**

- в данных таблицы hackathon_metrics выявлено 35 студентов, проходивших обучение на более чем 4 курсах Практикума. При этом 1 из них прошел 6 курсов
- 13% студентов сомневаются в прогрессе уверенности в знаниях (при этом более низкий показатель value наблюдается у дата инженеров)
- показатель lp_avg_user гораздо выше на курсах системного аналитика и дата инженера
- количество студентов, обучающихся по схеме b2b ничтожно - всего 24 человека, судя по всему корпаративная культура российских компаний не включает оплату курсов Практикума
- подавляющее число студентов - 82 % - не отвечали на вопрос Порекомендовали бы вы Практикум. Для сбора статстики или расчета NPS необходимо сделать ответ на данный вопрос обязательным

**2 Взаимосвязи между характеристиками студентов**

- не выявлено корреляции между стоимостью курса, его длительностья и lp_avg_user, value
- подтверждена гипотеза - чем больше опыт студента в сфере анализа данных и IT, тем выше lp_avg_user. При равном количестве лет опыта, студенты, работующие в сфере анализа имеют более высокий показатель, чем те, кто работал в сфере IT
- подтверждена гипотеза - студенты, нашедшие работу за время обучения, обладают более высоким показателем lp_avg_user, чем студенты, находящиеся в активном поиске или не ищущие работу
- не подтверждена гипотеза - студенты, записавшиеся в Карьерный Трек до сдачи диплома оценивают свою уверенность в улучшении знаний гораздо выше, чем те кто планируют или не хотят

**3 Портрет студента**

- большинство студентов выбирают спциальность аналитик данных или дата сайнтист (причем стандартный курс, не буткэмп и плюс)

- студенты учатся по схемам b2c или b2g
- основная цель студентов - сменить работу
- подавляющее число студетов 97% не меняли способа оплаты, всего 6,3% меняли когорту (уходили в академ) и более 99% не меняли курс

**4 Улучшение выстраивания помощи студентам**

- в зависимости от профессионального опыта студенты заинтересованы по-разному в предлагаемых Карьерным треком сервисах. Тем не менее, на первом месте по количеству заинтересованных всегда остается резюме. Для более персонализированного подхода к сопровождению студентов для каждой категории (по опыту) необходимо переопределить приоритетность той или иной информации и возможно сделать различную последовательность прохождения карьерного трека (или дать возможность студенту самому выбирать в какой последовательности он будет проходить блоки трека) с более детальной информацией по интересующим студентов направлениях в зависимости от их опыта
- отметим, что люди у которых есть опыт работы заинтересованы в определение стратегии поиска работы. Студенты без опыта, могут не осознавать важность данного аспекта. Можем порекомендовать Я.Практикуму добавить в блок трудоустройства пункт с дополнительной информацией о стратегии, но для самостоятельного изучения.
- у системных аналитиков (как и у дата инженеров, но мы их не берем в расчет из-за маленькой выборке) процент студентов без опыта в 2 раза меньше, чем у аналитиков и сайнтистов. Существуют курсы в направлении анализа данных, которые требуют меньшего сопровождения в плане ресурсов команды Карьерного трека, поскольку на них учаться студенты, уже работавшие в сфере IT или анализа, и, соответственно, имеющие большие шансы найти работу. Это подтверждается фактом что 16% системных аналитиков нашли работу в процессе обучения (против 4-5 у аналитиков и сайнтистов) Прооцент студентов, не хотевших записываться в карьерный трек, также выше у системных аналитиков (13% - системный аналитик, 9% - дата аналитик, 6% -дата сайнтист)

**5 Рекомендации**

Представленные к анализу данные не совсем коррелируют с конечной целью исследования, а именно сформулировать на основе данных гипотезы по улучшению выстраивания помощи студентам в достижении их целей. Необходима дополнительная информация по студентам, вступившим в Карьерный Трек

- цель
- дата начала участия в карьерном трека
- дата выгрузки информации
- цель достигнута/не достигнута и в результате чего (помощь карьерного трека, устройство на работу через знакомых и тд)
- принимал ли участие в оргнизованных Карьерным треком мероприятиях и если да, то в каких и сколько раз в качестве кого зрителя/участника (необходим сбор статистики, если таковой отсутвует)
- обращался ли лично за помощью к команде карьерного трека
- участвовал ли в мероприятиях, организованных Мастерскими Практикума (прокачка хардов)
- выгрузка статистики из кабинета Акселерации (для анализа количества откликов, используемых сайтов и методов связи с рекрутерами)

- для участников, достигнувших цели, дополнительно проводить анкетирования для выявления какая именно помощь карьерного трека была наиболее полезной

# 1 Предобработка данных

In [ ]:
```
#@title
!pip install kaleido
```

In [2]:
```
# Загрузка библиотек
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import kendalltau
import numpy as np
import plotly.express as px
# Снимаем ограничения по ширине таблицы
pd.set_option('display.max_columns', None)
pd.options.display.max_colwidth = 100
from scipy.stats import kruskal
from scipy.stats import f_oneway
sns.set(rc = {'figure.figsize':(15,8)})
import kaleido
import plotly
import plotly.graph_objects as go
```

## 1.1 hackathon_metrics

In [3]:
```
# чтение файла с данными и сохранение в hackathon_metrics
hackathon_metrics= pd.read_csv('https://raw.githubusercontent.com/EkaterinaTerentyeva/data_an
```

In [4]:
```
# Рассмотрим на наличие явных дубликатов
print(f'Количество дубликатов = {hackathon_metrics.duplicated().sum()}')
# просмотр информации о таблице
hackathon_metrics.info()
```

```
Количество дубликатов = 0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 79117 entries, 0 to 79116
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         79117 non-null  int64  
 1   profession_name    79117 non-null  object 
 2   user_id            79117 non-null  int64  
 3   lp_avg_user        79117 non-null  float64
 4   statement_content  79117 non-null  object 
 5   value              79117 non-null  int64  
 6   question_title     14739 non-null  object 
 7   user_answer        14739 non-null  float64
dtypes: float64(2), int64(3), object(3)
memory usage: 4.8+ MB
```

In [5]:
```
#Для дальнейшей работы сгруппируем данные по profession_name и user_id. Выведем последнюю  lea
hackathon_metrics_ = hackathon_metrics.groupby(['profession_name', 'user_id']).agg({'lp_avg_u
```

In [6]:
```
# определим сколько уникальных значений в данных колонки user_id
print(f"Количество студентов - {hackathon_metrics['user_id'].nunique()}")
# определим количество вопросов, на одного студента
print(f"Среднее количество вопросов на студента - {int(len(hackathon_metrics)/hackathon_metri
```

```
Количество студентов - 9797
Среднее количество вопросов на студента - 8
```

**В таблице приведены данные опроса 9797 студентов, дубликатов не обнаружено, около 80% пропущенных значений в user_answer и question_title (в дальнейшем определим, будем ли работать с этими признаками)**

# 1.2 data_goals_answers_fin

In [7]:
```
# чтение файла с данными и сохранение в data
data = pd.read_excel('https://raw.githubusercontent.com/EkaterinaTerentyeva/data_analyst_port
```

In [8]:
```
# Рассмотрим на наличие явных дубликатов
print(f'Количество дубликатов = {data.duplicated().sum()}')
# определим сколько уникальных значений в данных колонки user_id
print(f"Количество студентов = {data['user_id'].nunique()}")
# просмотр информации о таблице
data.info()
```

```
Количество дубликатов = 0
Количество студентов = 3549
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43428 entries, 0 to 43427
Data columns (total 17 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         43428 non-null  int64  
 1   user_id            43428 non-null  int64  
 2   answer_date        43428 non-null  object 
 3   answer_id          43428 non-null  object 
 4   cohort             43428 non-null  object 
 5   course_name        43428 non-null  object 
 6   current_cohort     43428 non-null  object 
 7   current_segment    43223 non-null  object 
 8   lesson_name        43428 non-null  object 
 9   original_segment   43223 non-null  object 
 10  profession_name    43428 non-null  object 
 11  question_title     43428 non-null  object 
 12  question_type      43428 non-null  object 
 13  slide_position     43428 non-null  int64  
 14  statement_content  0 non-null      float64
 15  topic_name         43428 non-null  object 
 16  user_answer        43416 non-null  object 
dtypes: float64(1), int64(3), object(13)
memory usage: 5.6+ MB
```

In [9]:
```
#Посчитаем какое количество вопросов приходится на пользователя
print(f"Количество уникальных вопросов = {data['question_title'].nunique()}")
#Какая доля студентов отвечает на все вопросы
print(f"Доля студентов, давших ответы на 6 вопросов = {round((data.groupby('user_id')['questi
```

```
Количество уникальных вопросов = 6
Доля студентов, давших ответы на 6 вопросов = 0.98
```

**В таблице приведены данные опроса 3549 студентов, дубликатов не обнаружено, в дальнейшем будут оставлены только нужные для анализа признаки. Студент как правило отвечал один раз на вопрос (98%)**

# 2 Анализ данных

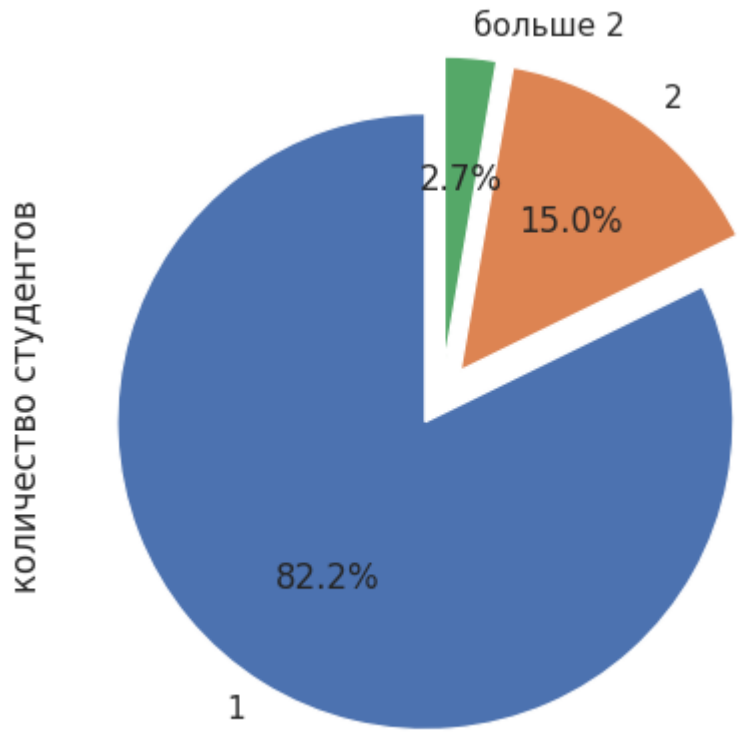
# 2.1 hackathon_metrics

## 2.1.1 Количество курсов

В данных явно есть любители Практикума - студенты, которые проходили несколько курсов (именно поэтому есть респонденты, ответичвшие более чем на 40 вопросов). Посмотрим на них

In [10]:
```
# Общее количество студетов в зависимости от количеству выбранных курсов
num_courses = hackathon_metrics.groupby('user_id')['profession_name'].nunique().sort_values(a
#Добавим столбец, который опрелеляет лояльность студентов к Я.Практикуму. Если студент купил
num_courses['лояльность_студентов'] = num_courses['количество курсов'].apply(lambda x: 'больш
#вывод результата
display(num_courses)
#Построение круговой диаграммы для результатов
num_courses_ = num_courses.groupby('лояльность_студентов')['количество студентов'].sum()
labels=num_courses['лояльность_студентов'].unique()
explode=(0.1,0.1,0.1)
num_courses_.plot.pie(y = 'лояльность_студентов', figsize=(5, 5), labels = labels, legend = F
```

| | количество курсов | количество студентов | лояльность_студентов |
|---|---|---|---|
| **0** | 1 | 8057 | 1 |
| **1** | 2 | 1472 | 2 |
| **2** | 3 | 233 | больше 2 |
| **3** | 4 | 26 | больше 2 |
| **4** | 5 | 8 | больше 2 |
| **5** | 6 | 1 | больше 2 |



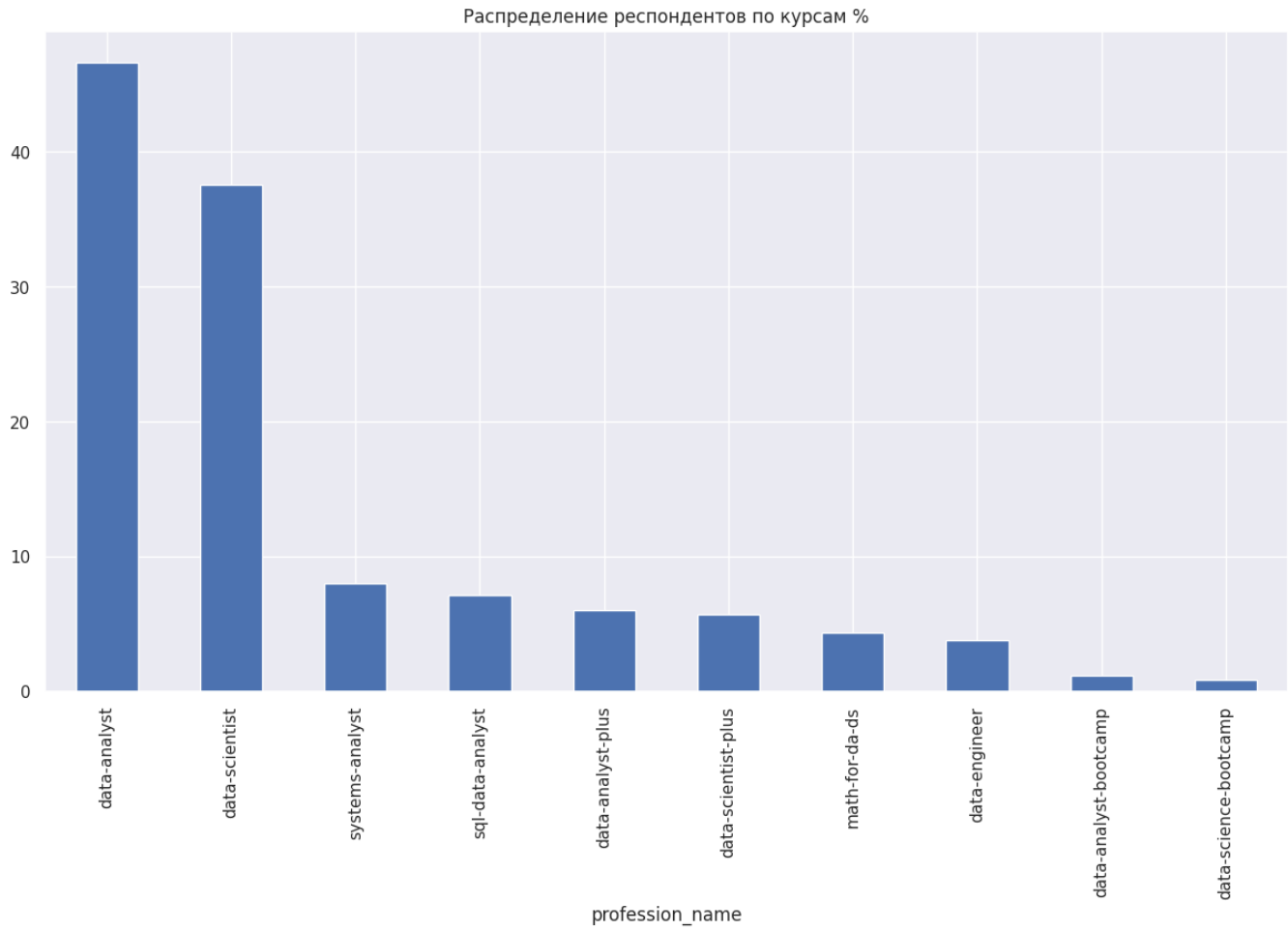

**У нас есть реальные поклонники Практикума - 35 человек, которые учились на не менее чем 4 курсах**

## 2.1.2 Популярные курсы и специальности

Рассмотрим какие курсы чаще выбирали наши респонденты

In [11]:
```
#Расчитаем соотношение количества студентов относительно курсов в процентах
pop_courses = (hackathon_metrics.groupby('profession_name')['user_id'].nunique()/hackathon_me
display(pop_courses)
#Построим гистограмму
pop_courses.sort_values(ascending = False).plot(kind = 'bar',  title = 'Распределение респонд
```

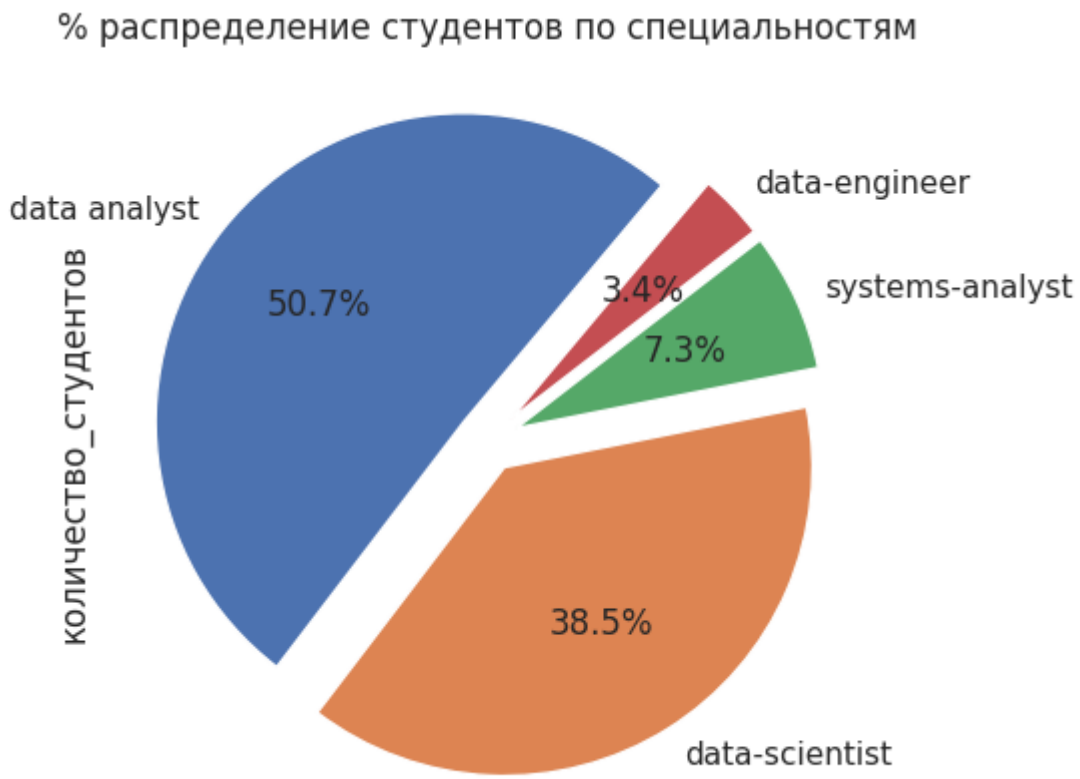
```
profession_name
data-analyst              46.61
data-analyst-bootcamp      1.15
data-analyst-plus          6.01
data-engineer              3.76
data-science-bootcamp      0.84
data-scientist            37.55
data-scientist-plus        5.65
math-for-da-ds             4.30
sql-data-analyst           7.10
systems-analyst            7.98
Name: user_id, dtype: float64
```




У нас по факту всего затронуто 4 специальности - аналитик данных, дата сайнтист,дата инженер, системный аналитик. И один курс - математика, который не является специальностью, а скорее дополняет основные курсы. Поэтому посмотрим распределение студентов именно по специальностям, а не по курсам

In [12]:
```
import numpy as np
#Функция для распределения курсов по специальностям
def get_speciality(cell):
    if 'data-analyst' in cell:
      return 'data analyst'
```

```
    elif 'scien' in cell:
      return 'data-scientist'
    elif 'math-for-da-ds' in cell or 'sql-data-analyst' in cell or 'datavis-and-bi-tools' in
      return np.nan
    else:
      return cell
#Добавим столбец со специальностями в hackathon_metrics
hackathon_metrics['speciality'] = hackathon_metrics['profession_name'].apply(get_speciality)
hackathon_metrics_['speciality'] = hackathon_metrics_['profession_name'].apply(get_speciality
# Соотношение количества студентов к специальностям
by_speciality = hackathon_metrics.groupby('speciality')['user_id'].nunique().sort_values(asce
#Построение круговой диаграммы для результатов
labels=by_speciality['speciality'].unique()
explode=(0.1,0.1,0.1, 0.1)
by_speciality.plot.pie(y = 'количество_студентов', title="% распределение студентов по специа
```




**Наибольшее число респондентов выбрали курсы анализа данных. Системный аналитик и дата инженер - сравнительно новые курсы Практикума, поэтому неудивительно, что для них наблюдается не так много респондентов**

## 2.1.3 Уверенность в повышении знаний

Рассмотрим среднюю оценку уверенности (value) по всем спринтам, после которых студент давал вопрос на ответ

In [13]:
```
# Средняя оценка в уверенности знаний относительно курса
data_value = round(hackathon_metrics_.groupby('profession_name')['value'].mean(), 2).sort_val
data_value
```

Out[13]:

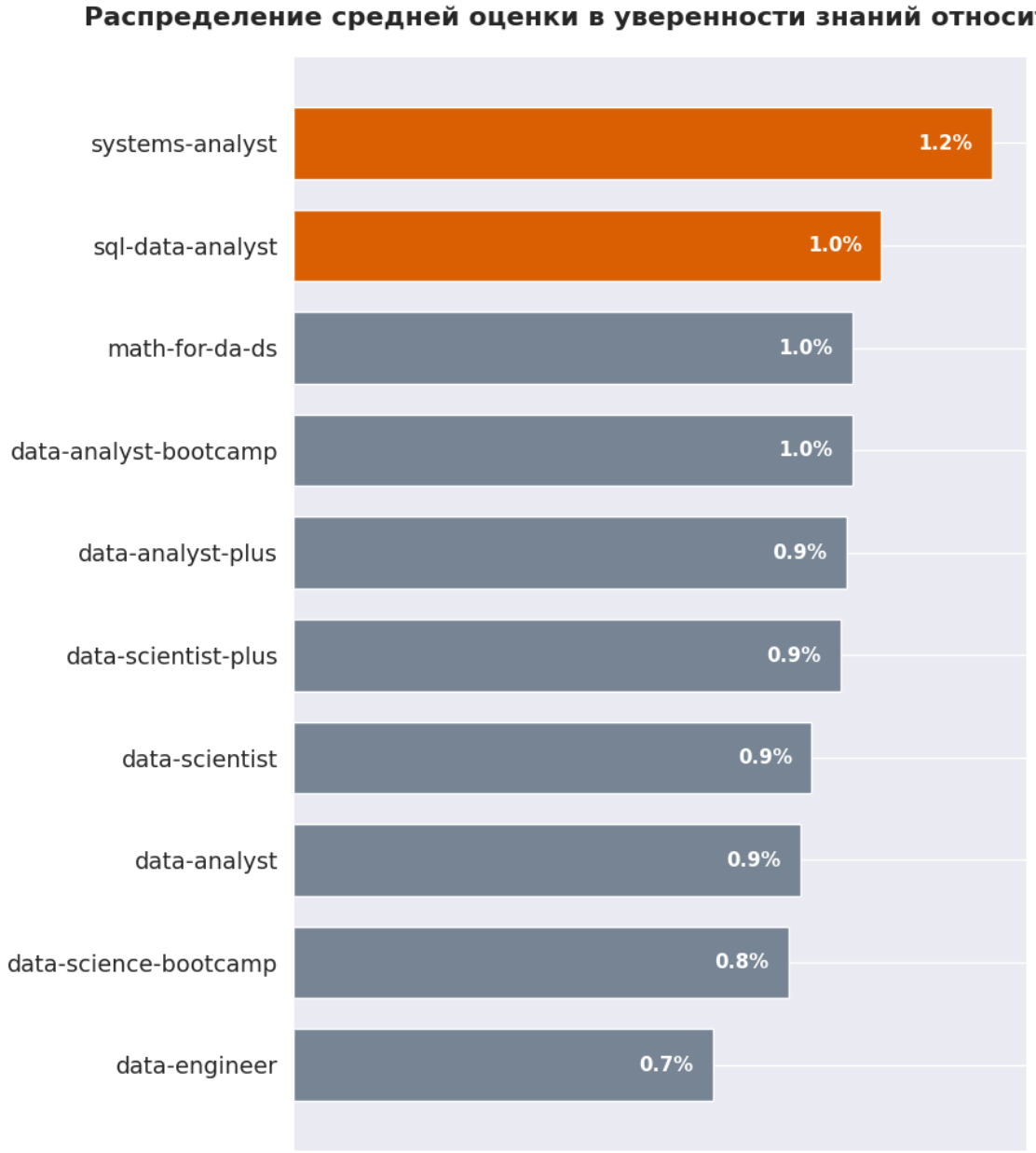
| | profession_name | value |
|---|---|---|
| **0** | systems-analyst | 1.20 |
| **1** | sql-data-analyst | 1.01 |
| **2** | data-analyst-bootcamp | 0.96 |
| **3** | math-for-da-ds | 0.96 |
| **4** | data-analyst-plus | 0.95 |
| **5** | data-scientist-plus | 0.94 |
| **6** | data-scientist | 0.89 |
| **7** | data-analyst | 0.87 |
| **8** | data-science-bootcamp | 0.85 |
| **9** | data-engineer | 0.72 |

In [14]:

```
#Построение графика для распределения средней оценки в уверенности знаний относительно курса
porosity_cutoff = 1
highlight_colour = '#d95f02'
non_highlight_colour = '#768493'

data_value['colours'] = data_value['value'].apply(lambda x: highlight_colour if x >= porosity
data_value = data_value.sort_values(by='value')
fig, ax = plt.subplots(figsize=(8,12))
bars = plt.barh(data_value['profession_name'], data_value['value'], height=0.7, color=data_va

ax.spines[['right', 'top', 'bottom']].set_visible(False)
ax.xaxis.set_visible(False)
ax.yaxis.set_tick_params(labelsize=14)
ax.bar_label(bars, padding=-45, color='white',
             fontsize=12, label_type='edge', fmt='%.1f%%',
            fontweight='bold')
ax.set_title('Распределение средней оценки в уверенности знаний относительно курса', fontsize
              fontweight='bold', pad=20)
plt.show()
```




Посмотрим есть ли респонденты, неуверенные в повышении своих знаниях

In [15]:
```
#Отсортируем студентов, которые не уверенны в повышении свих знаний value < 0
not_improuved = hackathon_metrics_.groupby(['profession_name', 'user_id'])['value'].mean().re
#Вывод результатов
print(f"Доля студентов, неуверенных в прогрессе своих знаний = {round(len(not_improuved)/hack
```

```
Доля студентов, неуверенных в прогрессе своих знаний = 0.13
```

**Дата инженеры имеют самый низкую оценку повышения своих знаний. Стоит отметить, что данный курс не подходит для студентов без опыта. Студенты,приступая к курсу должны иметь опыт работы с програмами. Стоит изучить закономерности между оценкой знаний, опытом и конечной целью студента на курсе**

**Около 13% студентов сомневаются в прогрессе уверенности в знаниях**

## 2.1.4 Анализ пропущенных значений

In [16]:
```
#Рассмотрим какое количество студентов ответило на вопрос "Какова вероятность, что вы пореком
print(f"Процент пропусков = {hackathon_metrics['user_answer'].isna().mean():.2%}")
print(f"Процент студентов, никогда не отвечавших на вопрос = {(hackathon_metrics.groupby('use
```

```
Процент пропусков = 81.37%
Процент студентов, никогда не отвечавших на вопрос = 86.35%
```

**Этот столбец абсолютно неинформативный, поскольку в нем содержится 81% пропусков (при этом 86% студентов вообще никогда не отвечали на этот вопрос). Подавляющее большинство студентов просто не считают нужным на него отвечать. Какаие-либо выводы по этому столбцу делать некорректно, единственное что мы можем сказать, что студенты не отвечают на этот вопрос и что для сбора статистику нужно сделать это поле обязательным для заполнения.**

## 2.1.5 Анализ успеваемости (learning performance)

Мы выбираем за показатель средний показатель - у нас нет даты в данных, поэтому мы не можем отследить прогресс по данному показателю по прохождению курса

In [17]:
```
#Рассчитаем средний learning performance для каждого курса
display(round(hackathon_metrics_.groupby('profession_name')['lp_avg_user'].mean(), 2).sort_va
#Рассчитаем средний learning performance для каждой специальности
display(round(hackathon_metrics_.groupby('speciality')['lp_avg_user'].mean(), 2).sort_values(
```

| | profession_name | lp_avg_user |
|---|---|---|
| **0** | systems-analyst | 0.89 |
| **1** | sql-data-analyst | 0.86 |
| **2** | data-engineer | 0.81 |
| **3** | data-analyst-bootcamp | 0.76 |
| **4** | data-science-bootcamp | 0.75 |
| **5** | data-scientist-plus | 0.74 |
| **6** | data-analyst-plus | 0.72 |
| **7** | data-scientist | 0.70 |
| **8** | data-analyst | 0.69 |
| **9** | math-for-da-ds | 0.66 |

| | speciality | lp_avg_user |
|---|---|---|
| **0** | systems-analyst | 0.89 |
| **1** | data-engineer | 0.81 |
| **2** | data analyst | 0.71 |
| **3** | data-scientist | 0.71 |

**Очень интересно, что на курсах системного аналитика и дата инженера этот показатель наиболее высокий - либо задания легче для выполнения, либо на эти курсы идут более опытные студенты или студенты с уже прокаченными хард скиллами (либо на этих курсах адекватный тренажер)**

## 2.1.6 Дополнительное исследование - зависимость показателей студентов от стоимости и продолжительности курса

**Дополним таблицу данными по продолжительности курса и его стоимости**

In [18]:
```
#Дополнительная информация взята с сайта Я.Практикума
#Цена за курс
price = [41000, 96000, 112000, 128000, 168000, 30000, 170000, 102400, 228000, 95000]
#Срок обучения
duration_month = [3, 6, 8, 4, 12, 4, 5, 8, 16, 6]
#Создание df
additional_info = pd.DataFrame()
additional_info['profession_name'] = hackathon_metrics['profession_name'].unique()
additional_info['price'] = price
additional_info['duration'] = duration_month
```

In [19]:
```
full_data = hackathon_metrics_.merge(additional_info, on = 'profession_name')
full_data[['value', 'lp_avg_user', 'price', 'duration']].corr()
```

Out[19]:

| | value | lp_avg_user | price | duration |
|---|---|---|---|---|
| **value** | 1.000000 | 0.075606 | -0.004382 | 0.010940 |
| **lp_avg_user** | 0.075606 | 1.000000 | -0.039972 | -0.012523 |
| **price** | -0.004382 | -0.039972 | 1.000000 | 0.934716 |
| **duration** | 0.010940 | -0.012523 | 0.934716 | 1.000000 |

In [20]:
```
# Расчет корреляции для полученных данных
full_data = hackathon_metrics.groupby(['user_id', 'profession_name'])[['value', 'lp_avg_user'
full_data[['value', 'lp_avg_user', 'price', 'duration']].corr()
```

Out[20]:

| | value | lp_avg_user | price | duration |
|---|---|---|---|---|
| **value** | 1.000000 | 0.079802 | -0.004382 | 0.010940 |
| **lp_avg_user** | 0.079802 | 1.000000 | -0.049575 | -0.020942 |
| **price** | -0.004382 | -0.049575 | 1.000000 | 0.934716 |
| **duration** | 0.010940 | -0.020942 | 0.934716 | 1.000000 |

**Никакой взаимосвязи между уверенностью в знаниях, а также средним learning perfomance и длительностью и стоимостью курса не найдено**

**ВЫВОД**

- Количество студентов - 9797. Среднее количество вопросов на студента - 8;
- У нас есть реальные поклонники Я.Практикума - 35 человек, которые учились на не менее чем 4 курсах;
- Наиболее востребована профессия аналитика данных;
- Наименее уверены в прогрессе своих знаний дата инженеры, при этом 13% студентов сомневаются что от спринта к спринту они более уверены в своих знаниях;
- Подавляющее число студентов - 82 % - не отвечали на вопрос "Порекомендовали бы вы Практикум". Для сбора статстики или расчета NPS необходимо сделать ответ на данный вопрос обязательным;
- Показатель learning performance на курсах системного аналитика и дата инженера наиболее высокий;
- Нет взаимосвязи между уверенностью в знаниях и learning perfomance со стоимостью и длительностью курса;

# 2.2 data_goals_answers_fin

## 2.2.1 Портрет студента

В качестве портрета студента мы выбираем следующие признаки:

- из какого сегмента был/стал студент;
- по какой специальности студент проходит обучение;
- какая у него конечная цель;
- какой опыт он имеет.

В данных опять-таки присутствуют курсы, которые не связаны с получением специальности, а скорее всего идут как дополнение к основным - sql-data-analyst, datavis-and-bi-tools, math-for-da-ds. Мы не будем их учитывать и сосредоточимся только на курсах, обучающих специальности. Кроме того встречаются ответы в разном стиле

In [21]:
```
#Уникальные значения в колонке profession_name
data['profession_name'].unique()
```

Out[21]:
```
array(['data-analyst', 'data-scientist', 'data-analyst-plus',
       'systems-analyst', 'data-scientist-plus', 'sql-data-analyst',
       'data-engineer', 'datavis-and-bi-tools', 'data-science-bootcamp',
       'math-for-da-ds', 'data-analyst-bootcamp'], dtype=object)
```

In [22]:
```
#Определим какие варианты ответов есть для вопроса "В зависимости от опыта работы вам может п
#от команды сопровождения и трудоустройства. Для нас очень важен честный ответ и понимание ва
data.query('question_title == "В зависимости от опыта работы вам может понадобиться разный ви
```

Out[22]:
```
array(['Нет опыта работы в IT и в направлении Анализа данных.',
       'От 1 года опыта работы в другом направлении IT.',
       'Более 3 лет опыта работы в направлении Анализа данных.',
       'Более 3 лет опыта работы аналитиком.',
       'От 1 до 3 лет опыта работы направлении Анализа данных.',
       'Менее 1 года опыта работы в другом направлении IT.',
       'От 1 до 3 лет опыта работы аналитиком.',
       'Нет опыта работы аналитиком и в IT.',
       'Менее года опыта работы в направлении Анализа данных.',
       'Нет опыта работы аналитиков и в IT.',
       'Менее года опыта работы аналитиком.'], dtype=object)
```

In [23]:
```
#Похожие варианты ответов объединим вместе
data['user_answer'] = (data['user_answer'].replace('Нет опыта работы аналитиком и в IT.','Нет
                                                  .replace('Нет опыта работы аналитиков и в IT.',
                                                  .replace('Более 3 лет опыта работы в направлени
                                                  .replace('От 1 до 3 лет опыта работы направлени
                                                  .replace('Менее года опыта работы в направлении
```

In [24]:
```
data['speciality'] = data['profession_name'].apply(get_speciality)
```

In [25]:
```
#Построим график, который отображает зависимость между тем на каком курсе студент из какого с
sankey = (data
          .query('question_title == "Бывает, что во время обучения меняется его цель. Наприме
          .groupby(['user_id', 'speciality'])['current_segment', 'user_answer']
          .first().reset_index())
all = sankey.groupby(['current_segment', 'speciality', 'user_answer'])['user_id'].count().res

categories = ['current_segment','speciality',   'user_answer']

newDf = pd.DataFrame()
for i in range(len(categories)-1):
    tempDf = all[[categories[i],categories[i+1],'user_id']]
    tempDf.columns = ['source','target','count']
    newDf = pd.concat([newDf,tempDf])
newDf = newDf.groupby(['source','target']).agg({'count':'sum'}).reset_index()

label_list = list(np.unique(all[categories].values))
source = newDf['source'].apply(lambda x: label_list.index(x))
target = newDf['target'].apply(lambda x: label_list.index(x))
count = newDf['count']
fig = go.Figure(data=[go.Sankey(
    node = {"label": label_list},
    link = {"source": source, "target": target, "value": count}
    )])
fig.show(renderer="colab");
```

```
<ipython-input-25-27f6848a85e3>:2: FutureWarning: Indexing with multiple keys (implicitly con
verted to a tuple of keys) will be deprecated, use a list instead.
  sankey = (data
```

**Как мы видим, компании очень редко оплачивают учебу своих сотрудников, практически за половину студентов платит государство, причем судя по всему курс системного аналитика и курс дата инженера не включены в программу Цифровые профессии. Подавляющее большинство респондентов пришли на курсы чтобы сменить работу (при этом в равной степени и сайнтисты, и аналитики, и инженеры) У нас всего 14 дата инженеров, принявших участие в опросе - недостаточный объем выборки для того, чтобы сделать какие то выводы для целого курса Практикума И только 40 студентов хотят открыть свой бизнес**

Также рассмотрим распределение по опыту

In [26]:
```
sankey_ = (data
           .query('question_title == "В зависимости от опыта работы вам может понадобиться раз
           .groupby(['user_id', 'speciality'])['user_answer']
           .first().reset_index())
all_ = sankey_.groupby(['speciality', 'user_answer'])['user_id'].count().reset_index()
#для лучшей визуализации укрупним категории по опыту
def get_cat(cell):
  if 'аналит' in cell:
    return 'опыт работы в анализе'
  elif 'Нет опыта работы' in cell:
    return 'без опыта'
  else: return 'опыт в IT'
all_['experience'] = all_['user_answer'].apply(get_cat)
all_ = all_.groupby(['speciality', 'experience'])['user_id'].sum().reset_index()
all_
```

Out[26]:

| | speciality | experience | user_id |
|---|---|---|---|
| **0** | data analyst | без опыта | 1347 |
| **1** | data analyst | опыт в IT | 282 |
| **2** | data analyst | опыт работы в анализе | 257 |
| **3** | data-engineer | без опыта | 2 |
| **4** | data-engineer | опыт в IT | 4 |
| **5** | data-engineer | опыт работы в анализе | 8 |
| **6** | data-scientist | без опыта | 871 |
| **7** | data-scientist | опыт в IT | 259 |
| **8** | data-scientist | опыт работы в анализе | 172 |
| **9** | systems-analyst | без опыта | 113 |
| **10** | systems-analyst | опыт в IT | 86 |
| **11** | systems-analyst | опыт работы в анализе | 114 |

In [27]:

```
#Построим график, который отображает зависимость между опытом и курсом
categories = ['source', 'target']
all_.columns = ['target', 'source', 'count']

label_list = list(np.unique(all_[categories].values))
source = all_['source'].apply(lambda x: label_list.index(x))
target = all_['target'].apply(lambda x: label_list.index(x))
count = all_['count']
fig = go.Figure(data=[go.Sankey(
    node = {"label": label_list},
    link = {"source": source, "target": target, "value": count}
    )])
fig.show(renderer="colab");
```

**Мы видим что 2/3 студентов курса системный аналитик уже имеют опыт в анализе или IT (этим и объясняется их лучший перфоманс)**

## 2.2.2 Определение признаков для сегментации

In [28]:
```
#Оставим в df колонки, необходимые для анализа
question = data[['user_id', 'question_title', 'user_answer', 'profession_name', 'speciality']
#Вывод результатов ответа, только для уникальных пользователей
answers = question.groupby(['user_id', 'question_title', 'profession_name', 'speciality'])['u
answers
```

Out[28]:

| | user_id | question_title | profession_name | speciality | user_answer |
|---|---|---|---|---|---|
| **0** | 3157 | Бывает, что во время обучения меняется его цель. Например, изначально вы не планировали менять р... | data-analyst | data analyst | Продвинуться по карьерной лестнице. |
| **1** | 3157 | В зависимости от опыта работы вам может понадобиться разный вид консультаций и помощи от команды... | data-analyst | data analyst | Нет опыта работы в IT и в направлении Анализа данных. |
| **2** | 3157 | Возможно вы нашли работу за время обучения? | data-analyst | data analyst | Да |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | определение профессиональной сферы |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | персональная карьерная консультация |
| **...** | ... | ... | ... | ... | ... |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | прохождение собеседований |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | как говорить про повышение |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | резюме |
| **20864** | 16535210 | Планируете ли вы записаться на Карьерный Трек? | data-analyst | data analyst | Да, уже записался |
| **20865** | 16535210 | Тут вы можете оставить свой комментарий, если не нашли подходящего варианта ответа. | data-analyst | data analyst | |

43040 rows × 5 columns

In [29]:
```
#Сохраним вопросы в отдельный лист для дальнейшей работы
list_answers = ['Бывает, что во время обучения меняется его цель. Например, изначально вы не
       'В зависимости от опыта работы вам может понадобиться разный вид консультаций и помощи
       'Возможно вы нашли работу за время обучения?',
       'Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут до
       'Планируете ли вы записаться на Карьерный Трек?']
```

Посмотрим на какие сегменты можно поделить студентов

- по специальности
- по оплате
- по тому переходили они из когорты в когорту или из одних курсов в другие
- по тому меняли ли они вообще курс на другой

In [30]:
```
# Доля студентов, которые не переходили в другую группу
(data['cohort'] == data['current_cohort']).mean().round(2)
```

Out[30]:
```
0.94
```

In [31]:
```
# Доля студентов, которые не меняли способ оплаты
(data['original_segment'] == data['current_segment']).mean().round(2)
```

Out[31]:
```
0.97
```

In [32]:
```
# Доля студентов, которые не меняли курс
data.groupby('user_id')['profession_name'].nunique().value_counts(normalize = True)
```

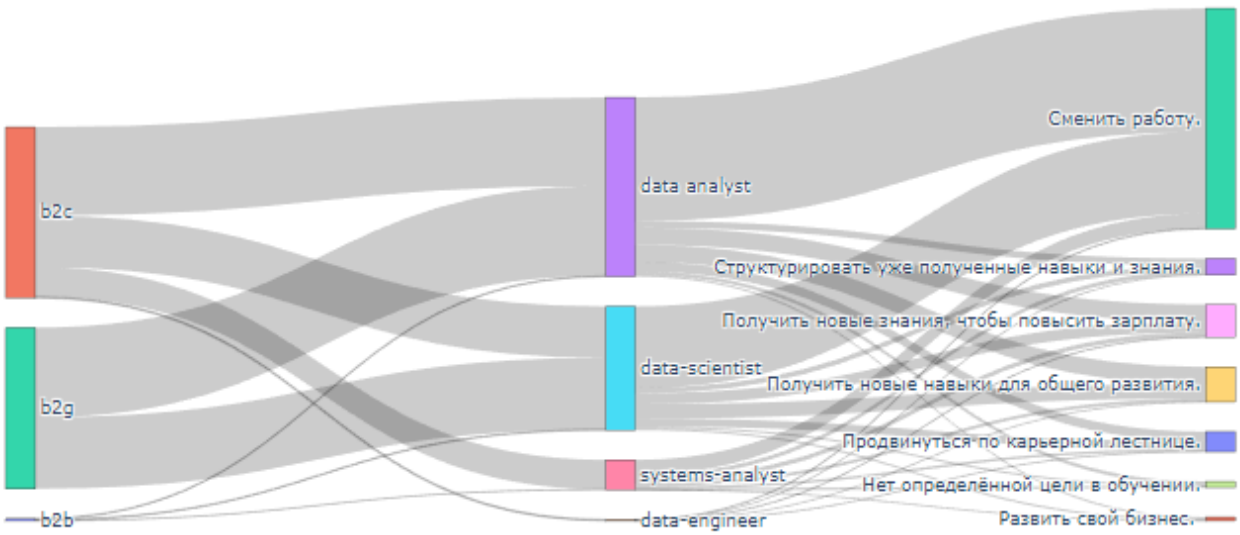
Out[32]:
```
1    0.997464
2    0.002254
3    0.000282
Name: profession_name, dtype: float64
```

Подавляющее число студетов не меняли способа оплаты, всего 6,3% меняли когорту (уходили в академ) и более 99% не меняли курс

**Таким основными признаками для построения сегментов будут являться выбранная специальность, профессиональный опыт и цель**

## 2.2.3 Определение востребованности тех или иных опций карьерного трека в зависимости от опыта

In [33]:
```
#Оставим данные которые, отвечают на конкретные вопросы: что интересует студента и их опыт
tab1 = answers.query('question_title     == "Для программы важно понять над чем вам нужно пор
tab2 =answers.query('question_title      == "В зависимости от опыта работы вам может понадоби
tab1
```



Out[33]:

| | user_id | question_title | profession_name | speciality | user_answer |
|---|---|---|---|---|---|
| **1** | 3157 | В зависимости от опыта работы вам может понадобиться разный вид консультаций и помощи от команды... | data-analyst | data analyst | Нет опыта работы в IT и в направлении Анализа данных. |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | определение профессиональной сферы |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | персональная карьерная консультация |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | устройство рынка труда |
| **3** | 3157 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | как и куда можно расти как специалисту |
| **...** | ... | ... | ... | ... | ... |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | сопроводительное письмо |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | как и куда можно расти как специалисту |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | прохождение собеседований |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | как говорить про повышение |
| **20863** | 16535210 | Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все ... | data-analyst | data analyst | резюме |

29158 rows × 5 columns

In [34]:

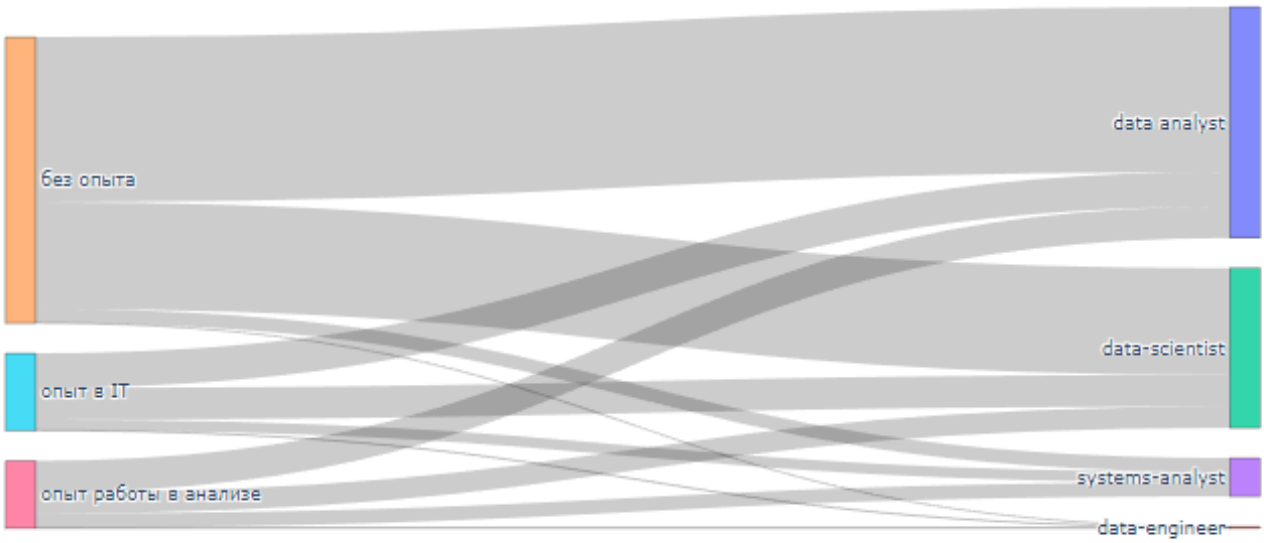
```
#Посчитаем количество пользователей относительно опыта
users = (answers.query('question_title == "В зависимости от опыта работы вам может понадобить
                .pivot_table(index = 'user_answer', values = 'user_id', aggfunc = 'count'))
users
```

Out[34]:

| user_answer | user_id |
|---|---|
| **Более 3 лет опыта работы аналитиком.** | 180 |
| **Менее 1 года опыта работы в другом направлении IT.** | 141 |
| **Менее года опыта работы аналитиком.** | 152 |
| **Нет опыта работы в IT и в направлении Анализа данных.** | 2333 |
| **От 1 года опыта работы в другом направлении IT.** | 491 |
| **От 1 до 3 лет опыта работы аналитиком.** | 219 |

In [35]:
```
#Выведим таблицу с результатами
pivot_table = pd.pivot_table(tab1.merge(tab2, on='user_id'),
                             values='user_id',
                             columns='user_answer_x',
                             index='user_answer_y',
                             aggfunc='count',
                             fill_value=0)
pivot_table = pivot_table.iloc[:, 6:]
res = pivot_table.join(users)
print('Процент студентов в зависимости от опыта, заинтересованных в той или иной опции Карьер
((res.div(res['user_id'], axis = 0)*100).astype('int')).drop('user_id', axis =1).style.backgr
```

```
Процент студентов в зависимости от опыта, заинтересованных в той или иной опции Карьерного Тр
ека
```



Out[35]:

| user_answer_y | как говорить про повышение | как и куда можно расти как специалисту | не думаю, что вы можете мне с чем-то помочь | определение профессиональной сферы | определение стратегии поиска работы | оформление портфолио | оценк трудоус |
|---|---|---|---|---|---|---|---|
| Более 3 лет опыта работы аналитиком. | 37 | 41 | 8 | 22 | 52 | 51 | |
| Менее 1 года опыта работы в другом направлении IT. | 45 | 53 | 10 | 46 | 71 | 73 | |
| Менее года опыта работы аналитиком. | 43 | 44 | 5 | 31 | 56 | 73 | |
| Нет опыта работы в IT и в направлении Анализа данных. | 32 | 50 | 4 | 52 | 69 | 79 | |
| От 1 года опыта работы в другом направлении IT. | 38 | 47 | 6 | 40 | 58 | 65 | |
| От 1 до 3 лет опыта работы аналитиком | 39 | 44 | 6 | 32 | 48 | 54 | |

**Мы видим, что абсолютно все студенты заинтересованы в первую очередь заинтересованы в составлении резюме и портфолио, далее их предпочтения меняются в зависимости от опыта**

**Отметит, что люди у которых есть опыт работы заинтересованы в определение стратегии поиска работы. Студенты без опыта, могут не осознавать важность данного аспекта.**

**Так же заострим внимание, что в курсах с углубленным изучением есть такая возможность как проработка стратегий поиска работы для студентов, а в базовом нет, но потребность как мы видим существует. Можем порекомендовать Я.Практикуму добавить в блок трудоустройства пункт с дополнительной информацией о стратегии, но для самостоятельного изучения.**

## 2.2.4 Ответы на вопросы в зависимости от выбранной специальности

In [36]:
```
#Цикл для вывода ответов на вопросы, опыта, активности  относительно специальности
for question in list_answers:
    print(question)
    display((answers.query('question_title == @question')
                    .pivot_table(index = 'speciality', columns = 'user_answer', values = 'use
                    .fillna(0).div(answers.groupby('speciality')['user_id'].nunique(), axis=0
                    .astype('int')
                    .style.background_gradient(axis =1))
    print('*'*200)
```

Бывает, что во время обучения меняется его цель. Например, изначально вы не планировали менять работу, но влюбились в профессию. Может, произошли жизненные изменения или вам сложно определить цель. Чтобы мы поняли, как помочь, отметьте подходящее утверждение:

| user_answer | Нет определённой цели в обучении. | Получить новые знания, чтобы повысить зарплату. | Получить новые навыки для общего развития. | Продвинуться по карьерной лестнице. | Развить свой бизнес. | Сменить работу. | Структурировать уже полученные навыки и знания. |
|---|---|---|---|---|---|---|---|
| speciality | | | | | | | |
| data analyst | 1 | 9 | 9 | 5 | 1 | 68 | 3 |
| data-engineer | 0 | 28 | 7 | 14 | 0 | 42 | 7 |
| data-scientist | 1 | 8 | 11 | 6 | 1 | 65 | 4 |
| systems-analyst | 1 | 15 | 10 | 8 | 0 | 49 | 13 |

**********

В зависимости от опыта работы вам может понадобиться разный вид консультаций и помощи от команды сопровождения и трудоустройства. Для нас очень важен честный ответ и понимание вашего бэкграунда.

| user_answer | Более 3 лет опыта работы аналитиком. | Менее 1 года опыта работы в другом направлении IT. | Менее года опыта работы аналитиком. | Нет опыта работы в IT и в направлении Анализа данных. | От 1 года опыта работы в другом направлении IT. | От 1 до 3 лет опыта работы аналитиком. |
|---|---|---|---|---|---|---|
| speciality | | | | | | |
| data analyst | 3 | 3 | 4 | 71 | 11 | 5 |
| data-engineer | 14 | 7 | 7 | 14 | 21 | 35 |
| data-scientist | 5 | 4 | 3 | 66 | 15 | 4 |
| systems-analyst | 10 | 5 | 8 | 36 | 22 | 17 |

**********

Возможно вы нашли работу за время обучения?

| user_answer | В активном поиске | Да | Нет |
|---|---|---|---|
| speciality | | | |
| data analyst | 16 | 4 | 76 |
| data-engineer | 35 | 14 | 35 |
| data-scientist | 15 | 4 | 78 |
| systems-analyst | 14 | 16 | 68 |

**********
**********
Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все опции.

| user_answer | как говорить про повышение | как и куда можно расти как специалисту | не думаю, что вы можете мне с чем-то помочь | определение профессиональной сферы | определение стратегии поиска работы | оформление портфолио | оценка трудоустр |
|---|---|---|---|---|---|---|---|
| speciality | | | | | | | |
| data analyst | 32 | 47 | 5 | 46 | 63 | 72 | |
| data-engineer | 50 | 42 | 7 | 28 | 50 | 57 | |
| data-scientist | 37 | 52 | 5 | 50 | 68 | 76 | |
| systems-analyst | 39 | 44 | 3 | 30 | 61 | 69 | |

**********
**********
Планируете ли вы записаться на Карьерный Трек?

| user_answer | Да, планирую записаться после диплома | Да, уже записался | Нет, не планирую |
|---|---|---|---|
| speciality | | | |
| data analyst | 52 | 37 | 9 |
| data-engineer | 57 | 21 | 21 |
| data-scientist | 58 | 35 | 5 |
| systems-analyst | 48 | 38 | 13 |

**********
**********

**Примечание - выборка для дата инженеров - 14 человек, приведенные результаты могут не отражать реальной ситуации, необходимо больше данных** Подавляющее большинство хочет сменить работу, при этом процент у аналитиков данных и сайнтистов гораздо выше. У системных аналитиков и дата инженеров больше опытных студентов. Дата инженерам вообще не интересно решение тестовых (если это вообще возможно)

## 2.3 Поиск закономерностей между успеваемостью/уверенностью студента и ответами на вопросы

### 2.3.1 Объединение данных

In [37]:
```
# Объеденим данные из таблиц hackathon_metrics и answers
united_data = hackathon_metrics_.merge(answers, on = 'user_id')
```

### 2.3.2 Анализ полученных значений по категориям

In [38]:
```
#Цикл для оценки закономерности между value,    lp_avg_user,    количество студентов и опытом
for answer in ['В зависимости от опыта работы вам может понадобиться разный вид консультаций
               'Возможно вы нашли работу за время обучения?',
               'Планируете ли вы записаться на Карьерный Трек?']:
    print(answer)
    display(united_data.query('question_title== @answer')
                       .groupby('user_answer').agg({'value':'mean', 'lp_avg_user':'mean', 'u
                       .sort_values(by = 'lp_avg_user', ascending = False)
                       .rename(columns = {'user_id':'количество студентов'}))
    print('*'*200)
```

```
В зависимости от опыта работы вам может понадобиться разный вид консультаций и помощи от кома
нды сопровождения и трудоустройства. Для нас очень важен честный ответ и понимание вашего бэк
граунда.
```

| user_answer | value | lp_avg_user | количество студентов |
|---|---|---|---|
| **Более 3 лет опыта работы аналитиком.** | 0.837679 | 0.753938 | 186 |
| **От 1 года опыта работы в другом направлении IT.** | 0.870360 | 0.748684 | 524 |
| **От 1 до 3 лет опыта работы аналитиком.** | 1.059130 | 0.742514 | 236 |
| **Менее года опыта работы аналитиком.** | 0.931349 | 0.741516 | 150 |
| **Менее 1 года опыта работы в другом направлении IT.** | 0.868939 | 0.714053 | 153 |
| **Нет опыта работы в IT и в направлении Анализа данных.** | 0.866603 | 0.704417 | 2543 |

```
****************************************************************************************************
****************************************************************************************************
**********
Возможно вы нашли работу за время обучения?
```

| user_answer | value | lp_avg_user | количество студентов |
|---|---|---|---|
| **Да** | 0.942027 | 0.757477 | 215 |
| **Нет** | 0.840550 | 0.717698 | 2894 |
| **В активном поиске** | 1.042873 | 0.704551 | 635 |

```
****************************************************************************************************
****************************************************************************************************
**********
Планируете ли вы записаться на Карьерный Трек?
```

| user_answer | value | lp_avg_user | количество студентов |
|---|---|---|---|
| **Да, уже записался** | 1.001919 | 0.721608 | 1443 |
| **Нет, не планирую** | 0.795894 | 0.721028 | 282 |
| **Да, планирую записаться после диплома** | 0.805524 | 0.713993 | 2046 |

**********
**********
**********

**Что здесь интересного**

- перфоманс lp_avg_user полностью соответствует опыту студента - большее количество лет опыта соответствует более высокому показателю
- люди, нашедшие работу за время обучения, обладают более высоким показателем lp_avg_user, чем студенты, находящиеся в активном поиске или не ищущие работу
- студенты, записавшиеся в Карьерный Трек до сдачи диплома оценивают свою уверенность в улучшении знаний гораздо выше, чем те кто планируют или не хотят при это их perfomance показатель не отличается от остальных

### 2.3.3 Проверка гипотез

**Гипотеза 1 - lp_avg_user различен у студентов с разным опытом и тем выше, чем больше опыт студента как в сфере анализа данных, так и в сфере IT. При этом lp_avg_user выше у студентов, работавших в сфере анализа данных по сравнению со студентами, работающими в сфере IT при аналогичном количестве лет опыта**

H0 - разницы в перфомансе между группами студентов по опыту нет

H1 - она есть

альфа = 0,05

Проверка осуществляется с помощью ANOVA и на всякий случай перепроверяется с помощью непараметрического теста kruskal

In [39]:
```
groups = []
for i in united_data.query('question_title == "В зависимости от опыта работы вам может понадо
    groups.append(united_data.query('user_answer == @i')['lp_avg_user'])

f_statistic, p_value_anova = f_oneway(groups[0], groups[1], groups[2], groups[3], groups[4],
h_statistic, p_value_kruskal = kruskal(groups[0], groups[1], groups[2], groups[3], groups[4],
print("p-значение ANOVA:", p_value_anova)
print("p-значение Kruskal:", p_value_kruskal)
```

```
p-значение ANOVA: 1.1671575862945173e-16
p-значение Kruskal: 2.02829725425213e-15
```

**Гипотеза 1 подтвердилась**

**Гипотеза 2 - студенты, нашедшие работу за время обучения, обладают более высоким показателем lp_avg_user, чем студенты, находящиеся в активном поиске или не ищущие работу**

H0 - разницы в перфомансе между группами студентов по результату поиска работы нет

H1 - она есть

альфа = 0,05

Проверка осуществляется с помощью ANOVA и на всякий случай перепроверяется с помощью непараметрического теста kruskal

In [40]:
```
groups = []
for i in united_data.query('question_title == "Возможно вы нашли работу за время обучения?"')
    groups.append(united_data.query('user_answer == @i')['lp_avg_user'])

f_statistic, p_value_anova = f_oneway(groups[0], groups[1], groups[2])
h_statistic, p_value_kruskal = kruskal(groups[0], groups[1], groups[2])
print("p-значение ANOVA:", p_value_anova)
print("p-значение Kruskal:", p_value_kruskal)
```

```
p-значение ANOVA: 1.5702888660629418e-06
p-значение Kruskal: 1.4584265508382501e-06
```

**Гипотеза 2 подтвердилась**

**Гипотеза 3 - студенты, записавшиеся в Карьерный Трек до сдачи диплома оценивают свою уверенность в улучшении знаний гораздо выше, чем те кто планируют или не хотят при это их perfomance показатель не отличается от остальных**

H0 - разницы в уверенности в знаниях между группами студентов по началу прохождения Карьерного Трека нет

H1 - она есть

альфа = 0,05

Проверка осуществляется с помощью ANOVA и на всякий случай перепроверяется с помощью непараметрического теста kruskal

In [41]:
```
groups = []
for i in united_data.query('question_title == "Планируете ли вы записаться на Карьерный Трек?
    groups.append(united_data.query('user_answer == @i')['lp_avg_user'])

f_statistic, p_value_anova = f_oneway(groups[0], groups[1], groups[2])
h_statistic, p_value_kruskal = kruskal(groups[0], groups[1], groups[2])

print("p-значение ANOVA:", p_value_anova)
print("p-значение Kruskal:", p_value_kruskal)
```

```
p-значение ANOVA: 0.21359380133385208
p-значение Kruskal: 0.12818122444158933
```

**Гипотеза 3 не подтвердилась**

Мы подтвердили 1 и 2 выдвинутые гипотезы - для каждой из них уровень p-value меньше выбранного альфа (при этом мы рассчитывали p-value с помощью ANOVA и его непараметрического аналога на случай если распределение не отвечает нормальности и дисперсии групп неоднородны)