Цели обучения в Яндекс Практикум - результаты опроса студентов

Ссылки на материалы

1. **Аналитический проект** (Colab)
2. **Презентация** (Figma)
3. **Архив с графиками в .SVG**

План проекта

1. Введение

Отрасль и направления деятельности: EdTech, сервис-онлайн образования

О проекте: Создание и оформление отчёта целей обучения студентов Яндекс Практикума для презентации топ-менеджменту Яндекс Практикума.

Задачи 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.

2. 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

3. Загрузка данных

Импортируем библиотеки.

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\user\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Объявим функции.

Загрузим исходные данные.

```
array(['Моя уверенность в своих знаниях значительно повысилась по сравнению с пред
ыдущим спринтом'],
      dtype=object)
```

Выведем первые строки таблиц.

	user_id	answer_date	answer_id	cohort	course_name	current_cohort	current_segm
1	3157	2023-09-01 10:43:00	41ac1a75-1f3b-44bd-a2f2-346c3bdef7b3	data_cohort_121	Трудоустройство	data_cohort_121	b
2	3157	2023-09-01 10:43:00	4c9d62c8-beed-4cab-a48a-a7168dbf9fdf	data_cohort_121	Трудоустройство	data_cohort_121	b
3	3157	2023-09-01 10:42:00	fe97eac2-5e16-4e28-9aab-83669b4c5629	data_cohort_121	Трудоустройство	data_cohort_121	b
4	3157	2023-09-01 10:43:00	41ac1a75-1f3b-44bd-a2f2-346c3bdef7b3	data_cohort_121	Трудоустройство	data_cohort_121	b
5	3157	2023-09-01 10:43:00	41ac1a75-1f3b-44bd-a2f2-346c3bdef7b3	data_cohort_121	Трудоустройство	data_cohort_121	b

	profession_name	user_id	lp_avg_user	statement_content	value	question_title	user_answer
0	sql-data-analyst	14641026	0.860000	Моя уверенность в своих знаниях значительно по...	2	Какова вероятность, что вы порекомендуете Прак...	9.0
1	sql-data-analyst	14641026	0.930000	Моя уверенность в своих знаниях значительно по...	2	Какова вероятность, что вы порекомендуете Прак...	9.0
2	data-analyst	14881168	0.480000	Моя уверенность в своих знаниях значительно по...	1	NaN	NaN
3	data-analyst	14881168	0.426667	Моя уверенность в своих знаниях значительно по...	1	NaN	NaN
4	data-scientist	7855703	0.611111	Моя уверенность в своих знаниях значительно по...	2	NaN	NaN

Удостоверимся в соответствии типов данных ожидаемым.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 43428 entries, 1 to 43428
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   user_id            43428 non-null  int64
 1   answer_date        43428 non-null  datetime64[ns]
 2   answer_id          43428 non-null  object
 3   cohort             43428 non-null  object
 4   course_name        43428 non-null  object
 5   current_cohort     43428 non-null  object
 6   current_segment    43223 non-null  object
 7   lesson_name        43428 non-null  object
 8   original_segment   43223 non-null  object
 9   profession_name    43428 non-null  object
 10  question_title     43428 non-null  object
 11  question_type      43428 non-null  object
 12  slide_position     43428 non-null  int64
 13  statement_content  0 non-null      float64
 14  topic_name         43428 non-null  object
 15  user_answer        43416 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(12)
memory usage: 5.6+ MB
None
<class 'pandas.core.frame.DataFrame'>
Int64Index: 79117 entries, 0 to 79116
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   profession_name    79117 non-null  object
 1   user_id            79117 non-null  int64
 2   lp_avg_user        79117 non-null  float64
 3   statement_content  79117 non-null  object
 4   value              79117 non-null  int64
 5   question_title     14739 non-null  object
 6   user_answer        14739 non-null  float64
dtypes: float64(2), int64(2), object(3)
memory usage: 4.8+ MB
None
```

Типы данных соответствуют ожидаемым.

Первичные проверки после загрузки проведены, приступаем к предобработке.

4. Предобработка данных

4.1 Пропуски и дубликаты в данных

Проверим данные на пропуски.

	na_sum	na_percent
statement_content	43428	100.000000
current_segment	205	0.470000
original_segment	205	0.470000
user_answer	12	0.030000
user_id	0	0.000000

answer_date	0	0.000000
answer_id	0	0.000000
cohort	0	0.000000
course_name	0	0.000000
current_cohort	0	0.000000
lesson_name	0	0.000000
profession_name	0	0.000000
question_title	0	0.000000
question_type	0	0.000000
slide_position	0	0.000000
topic_name	0	0.000000

	na_sum	na_percent
question_title	64378	81.370000
user_answer	64378	81.370000
profession_name	0	0.000000
user_id	0	0.000000
lp_avg_user	0	0.000000
statement_content	0	0.000000
value	0	0.000000

В датафрейме `df_answers` с ответами пользователей столбец `statement_content` пустой полностью, поскольку в этой таблице вопросы сформулированы в другом столбце, `question_title`.

У 205 пользователей не указан сегмент в столбцах `current_segment` и `original_segment`, то есть для этих пользователей неизвестно, кто оплачивал курс - сам студент, работодатель или оплата была по госпрограмме. Процент пустых значений менее 0.5%, на качество анализа наличие пустых значений здесь не повлияет.

В датафрейме `df_metrics` в 81.4% строк отсутствуют значения в столбцах `question_title` и `user_answer`, отвечающих соответственно за следующий вопрос и ответ на него.

```
'Какова вероятность, что вы порекомендуете Практикум своим друзьям по шкале от 0 д
о 10, где 10 — обязательно порекомендую, 0 — не порекомендую ни за что?'
```

Судя по всему, большая часть пользователей пропускает этот вопрос и не отвечает на него. Нужно учесть это в дальнейшем анализе при формировании методики оценки удовлетворённости студентов от обучения.

Проверим таблицы на наличие дубликатов в данных, начнём с `df_metrics`.

```
Full duplicates: 41817 (52.85%)
```

	duplicates	duplicates_percent
statement_content	79116	100.000000

question_title	79115	100.000000
value	79112	99.990000
profession_name	79107	99.990000
user_answer	79105	99.980000
lp_avg_user	75543	95.480000
user_id	69320	87.620000

В таблице **df_metrics** полные дубликаты могли появиться в результате ответа студента на вопрос после каждого спринта. Посмотрим, какие максимальные значения количества ответов на студента.

user_id	value	user_answer
13913671	64	64
2560890	60	60
15074343	60	60
14874784	55	0
3178885	54	54
13771402	52	0
909067	50	50
1785504	48	48
15150991	42	0
14693840	42	42

Проверим, есть ли студенты, проходящие обучение сразу по нескольким профессиям.

```
1    8057
2    1472
3     233
4      26
5       8
6       1
Name: profession_name, dtype: int64
```

Поскольку такие студенты есть, очистим датафрейм от дубликатов, оставив для каждого для каждой пары **user_id** - **profession_name** одну строку со средними `user_answer` и `value` и последним `lp_avg_user`. Также удалим столбцы с формулировками вопросов, а столбцы с ответами на них переименуем для ясности.

```
Full duplicates: 0 (0.0%)
```

	duplicates	duplicates_percent
confidence	11845	99.960000
profession_name	11840	99.920000
recom_rate	11838	99.900000
lp_avg_user	9191	77.560000

user_id	2053	17.320000

Дубликаты по отдельным полям ожидаемы, т.к. варианты значений в них строго определены и повторяются. Дубликаты в поле `user_id` вызваны тем, что один студент может учиться разным профессиям и в этих случаях мы сохранили для каждого такого студента по одной строке на каждую профессию.

Посмотрим на дубликаты в `df_answers`.

```
Full duplicates: 0 (0.0%)
```

	duplicates	duplicates_percent
course_name	43427	100.000000
lesson_name	43427	100.000000
statement_content	43427	100.000000
topic_name	43427	100.000000
slide_position	43426	100.000000
question_type	43425	99.990000
current_segment	43424	99.990000
original_segment	43424	99.990000
question_title	43422	99.990000
profession_name	43417	99.970000
cohort	43259	99.610000
current_cohort	43255	99.600000
user_answer	42325	97.460000
user_id	39879	91.830000
answer_date	35546	81.850000
answer_id	29334	67.550000

Полные дубликаты отсутствуют. Дубликаты в отдельных полях ожидаемы: большинство из них имеют строго определённые варианты значений и повторяются.

Для поля `user_id` дубликаты объясняются тем, что каждый студент отвечал множество вопросов, на каждый ответ - новая строка.

Дубликаты в поле `answer_id` могут объясняться наличием вопросов, где могло быть выбрано несколько ответов, на каждый из которых - своя строка.

4.2 Удаление лишних столбцов

Рассмотрим таблицу `df_answers` на предмет столбцов, которые можно удалить, т.к. они не важны с точки зрения нашего анализа.

Начнём со столбцов, в которых содержатся единственные значения признаков или вовсе пустые.

```
topic_name           1
statement_content    0
lesson_name          1
course_name          1
dtype: int64
```

Помимо вышеперечисленных столбцов, не содержащих нужных для анализа данных, есть столбец `slide_position` — исходя из описания данных, он содержит "страницу опроса" и для анализа не нужен.

Удалим ненужные столбцы из датафрейма.

Чтобы понять, нужен ли нам столбец `original_segment`, рассмотрим столбцы со значением текущего сегмента пользователя и его предыдущего сегмента - увидим, как менялись сегменты у студентов.

```
original_segment  current_segment
b2g               b2c                993
dtype: int64
```

Итак, во всех случаях, когда сегмент у студентов менялся, это была смена с сегмента b2g на b2c. Чтобы это стало понятнее в данных, заменим столбец `original_segment` на столбец `from_b2g`, где проставим значение `True` во всех случаях, когда сегмент менялся с b2g на b2c и `False`, когда не менялся. Сам столбец `original_segment` удалим.

Похожую операцию проделаем со столбцами `cohort` и `current_cohort`, сформировав столбец `cohort_changed` с признаком того, что когорта была изменена - возможно, эту категорию студентов будет интересно рассмотреть отдельно, а столбец `cohort` с конкретным указанием на изначальную когорту удалим.

4.3 Преобразование и соединение таблиц и *обработка* итоговой таблицы

Посмотрим, какие у нас вопросы, типы ответов и сколько вариантов ответов.

question_title	question_type	possible_answers
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:	radio	7
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.	radio	11
Возможно вы нашли работу за время обучения?	radio	3
Для программы важно понять над чем вам нужно поработать. Не переживайте, вам будут доступны все опции.	checkbox	13
Планируете ли вы записаться на Карьерный Трек?	radio	3

Тут вы можете оставить свой комментарий, если не нашли подходящего варианта ответа.	text	1067

Изменим таблицу так, чтобы колонки с вопросами и ответами из длинной формы преобразовать в широкую, при этом столбец `question_type` в новую таблицу не добавляем, т.к. он нам более не понадобится.

После такого преобразования в таблице должны были появиться полные дубликаты, которых раньше не было, проверим.

```
29334
```

Удалим полные дубликаты и проверим результат, также по ключевым полям.

```
Full duplicates: 0 (0.0%)
```

	duplicates	duplicates_percent
user_id	10545	74.820000
answer_id	0	0.000000

Учитывая, что мы уже преобразовали вопросы в широкую форму, дубликатов по `user_id` слишком много. По всей видимости, студент мог давать ответы на одни и те же вопросы в разные даты. Оставим по одному самому позднему ответу для каждой пары **user_id-profession_name**.

Посмотрим, сколько в итоге строк.

```
Full duplicates: 0 (0.0%)
```

	duplicates	duplicates_percent
user_id	10	0.280000
answer_id	0	0.000000

Соединим с таблицей `data` по паре **user_id - profession_name**.

	user_id	answer_date	answer_id	current_cohort	current_segment	profession_name	f
0	3157	2023-09-01 10:43:00	4c9d62c8-beed-4cab-a48a-a7168dbf9fdf	data_cohort_121	b2g	data-analyst	
1	5415	2023-08-03 13:40:00	d356d61d-ed3a-4e2f-8c20-0b3d687df655	data_cohort_119	b2g	data-analyst	
2	8199	2023-07-29 17:12:00	1d30aceb-2955-44ca-a4ff-5ea7562e1c8a	data_cohort_103	b2c	data-analyst	
3	8215	2023-07-25 21:39:00	d56b4e11-a9b3-49be-	ds_cohort_101	b2c	data-scientist	

			8e7e-00af6327e389				
4	10202	2023-07-18 14:46:00	3fd66b68-395c-48b3-9e52-188115ccef43	data_analyst_plus_cohort_31		b2g	data-analyst-plus

5 rows × 29 columns

Изменим для удобства порядок столбцов.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3559 entries, 0 to 3558
Data columns (total 29 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   answer_id                                     3559 non-null   object
 1   answer_date                                   3559 non-null   datetime64[ns]
 2   user_id                                       3559 non-null   int64
 3   profession_name                               3559 non-null   object
 4   current_cohort                                3559 non-null   object
 5   cohort_changed                                3559 non-null   bool
 6   current_segment                               3532 non-null   object
 7   from_b2g                                      3559 non-null   bool
 8   lp_avg_user                                   3146 non-null   float64
 9   confidence                                    3146 non-null   float64
 10  recom_rate                                    534 non-null    float64
 11  q1_goal                                       254 non-null    object
 12  q2_background                                 89 non-null     object
 13  q3_job_status                                 2161 non-null   object
 14  q4_career_track                               254 non-null    object
 15  q5_text_comment                               2161 non-null   object
 16  как говорить про повышение                    365 non-null    float64
 17  как и куда можно расти как специалисту        521 non-null    float64
 18  не думаю, что вы можете мне с чем-то помочь   46 non-null     float64
 19  определение профессиональной сферы            494 non-null    float64
 20  определение стратегии поиска работы           683 non-null    float64
 21  оформление портфолио                          789 non-null    float64
 22  оценка шансов на трудоустройство              687 non-null    float64
 23  персональная карьерная консультация           578 non-null    float64
 24  прохождение собеседований                     776 non-null    float64
 25  резюме                                        803 non-null    float64
 26  решение тестовых заданий                      751 non-null    float64
 27  сопроводительное письмо                       726 non-null    float64
 28  устройство рынка труда                        509 non-null    float64
dtypes: bool(2), datetime64[ns](1), float64(16), int64(1), object(9)
memory usage: 785.5+ KB
```

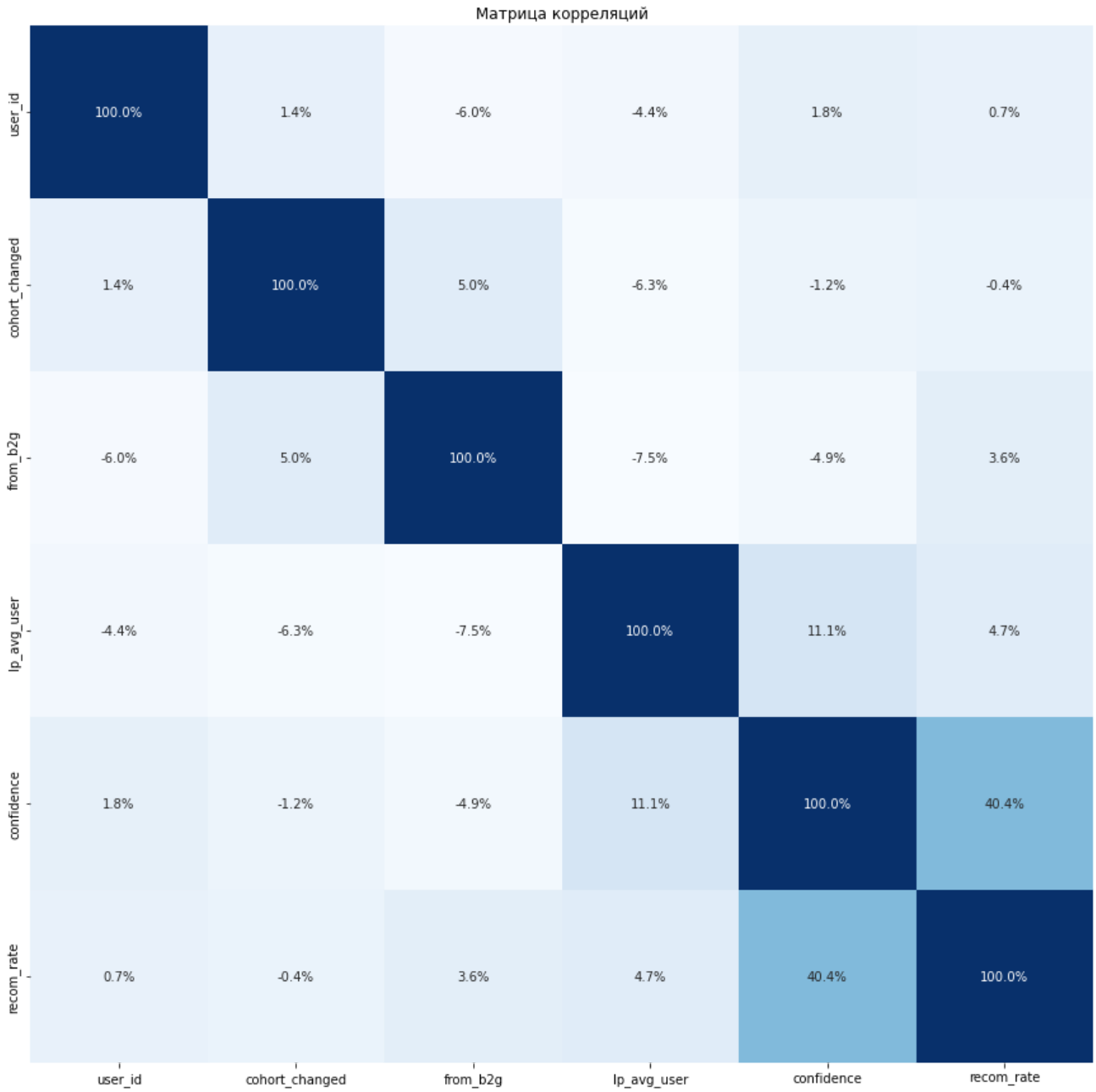
Выводы предобработки:

1. Студенты часто пропускают вопросы, не отвечая на них.
2. Студенты отвечают на вопросы регуляно, в этом случае мы в анализе учитываем последний ответ.
3. Один студент может одновременно обучаться нескольким профессиям, в этом случае мы учитываем последние ответы студента по каждой профессии.

5. Исследовательский анализ

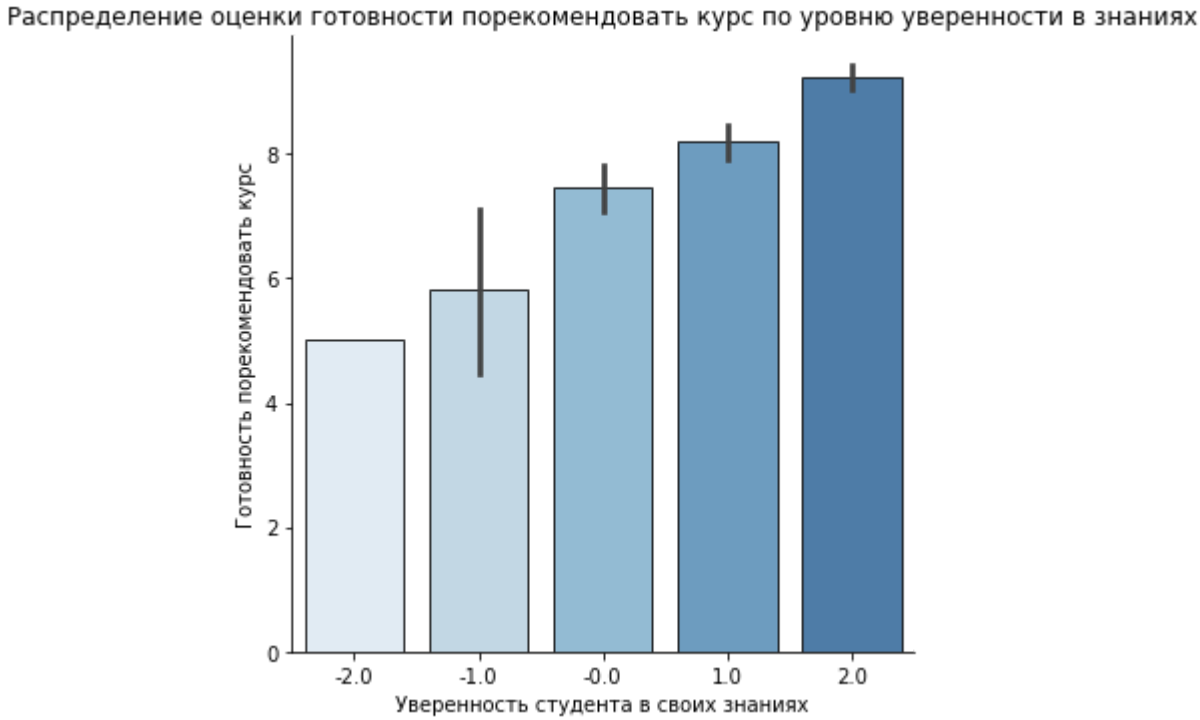
5.1 Матрица корреляций

Начнём исследовательский анализ с поиска корреляций между количественными значениями в данных.



Предположительно есть корреляция между двумя параметрами, посмотрим ближе:

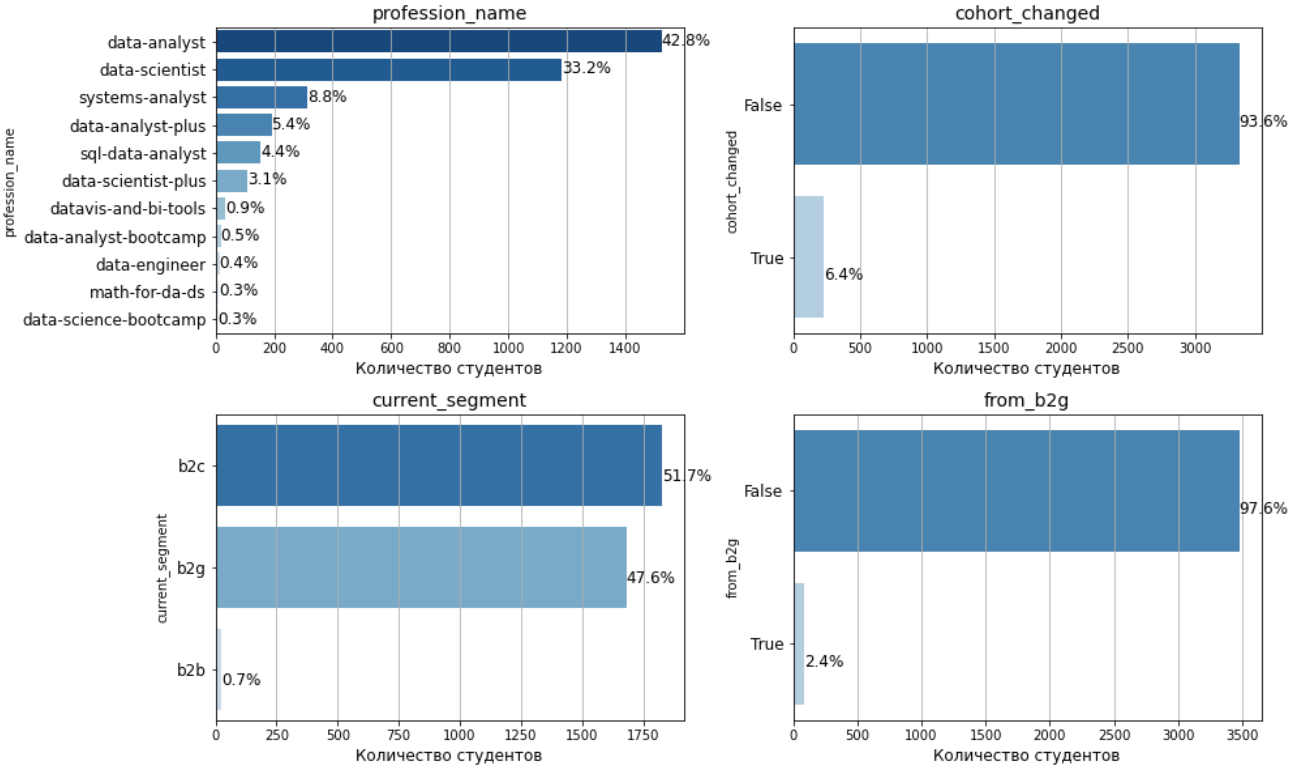
```
<Figure size 864x576 with 0 Axes>
```

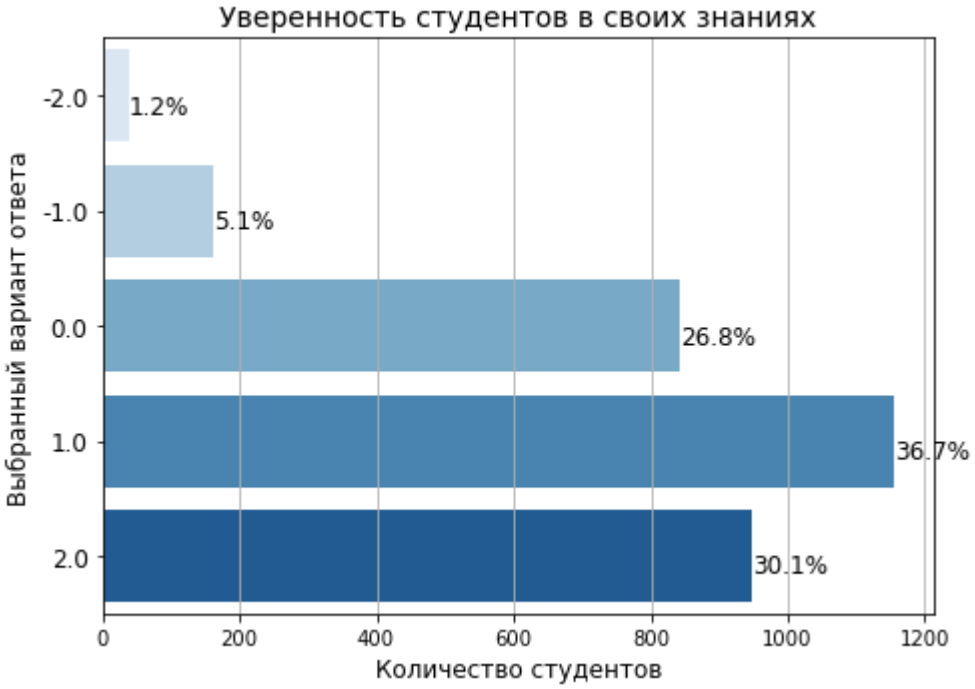


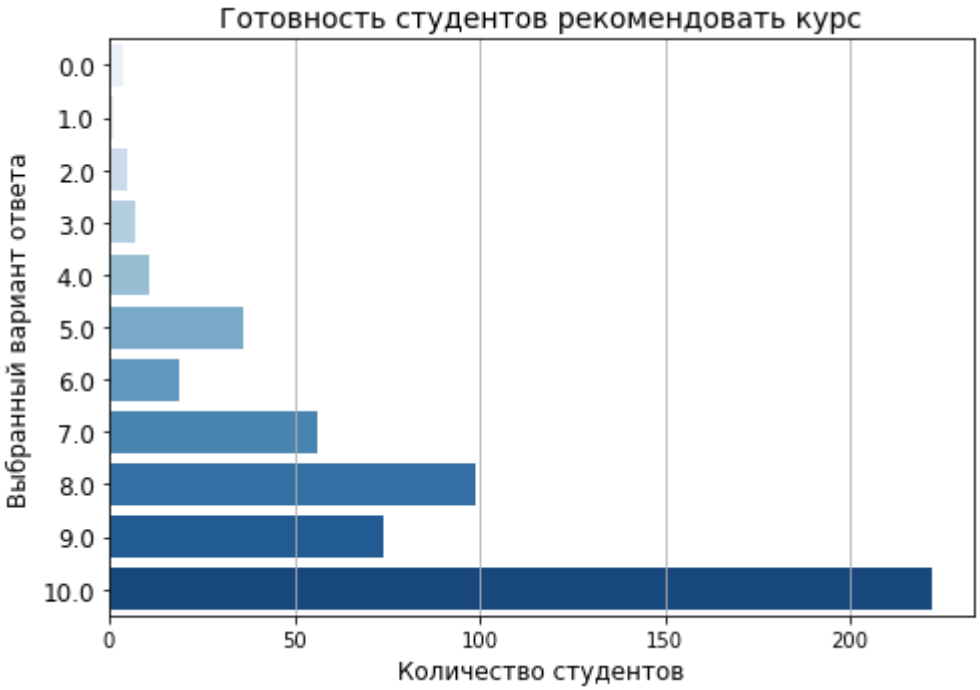
Средняя оценка готовности порекомендовать курс увеличивается с ростом уверенности студента в своих знаниях.

5.2 Основные признаки датасета

Визуализируем основные признаки датасета:

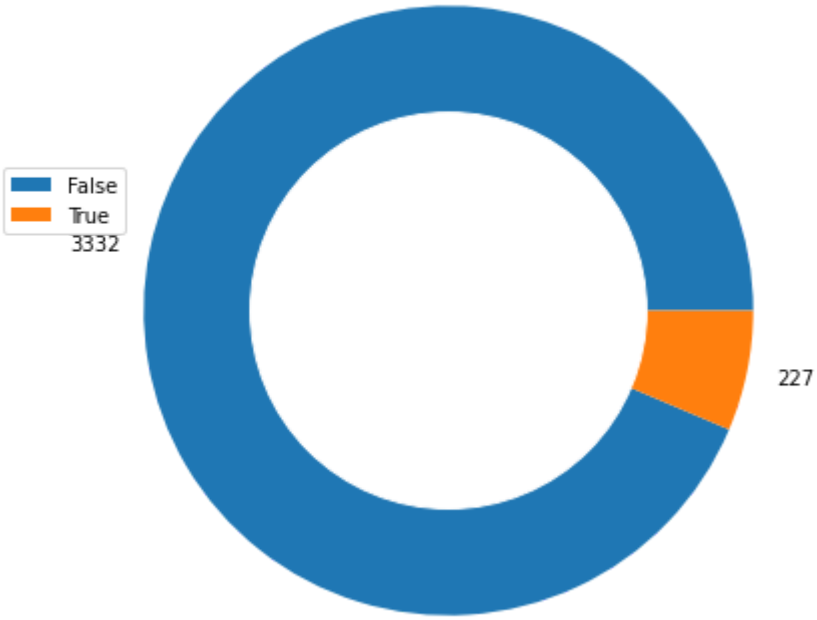




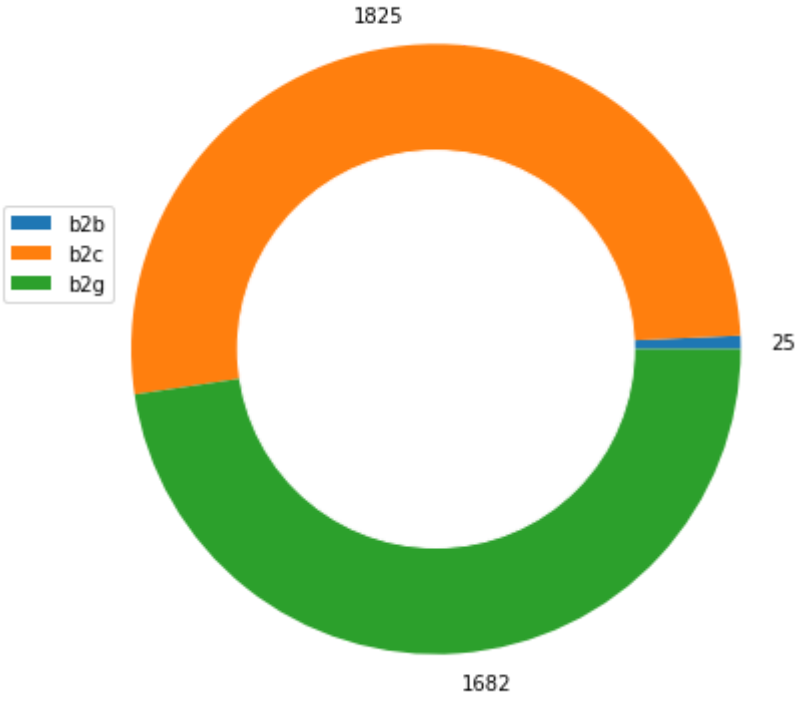


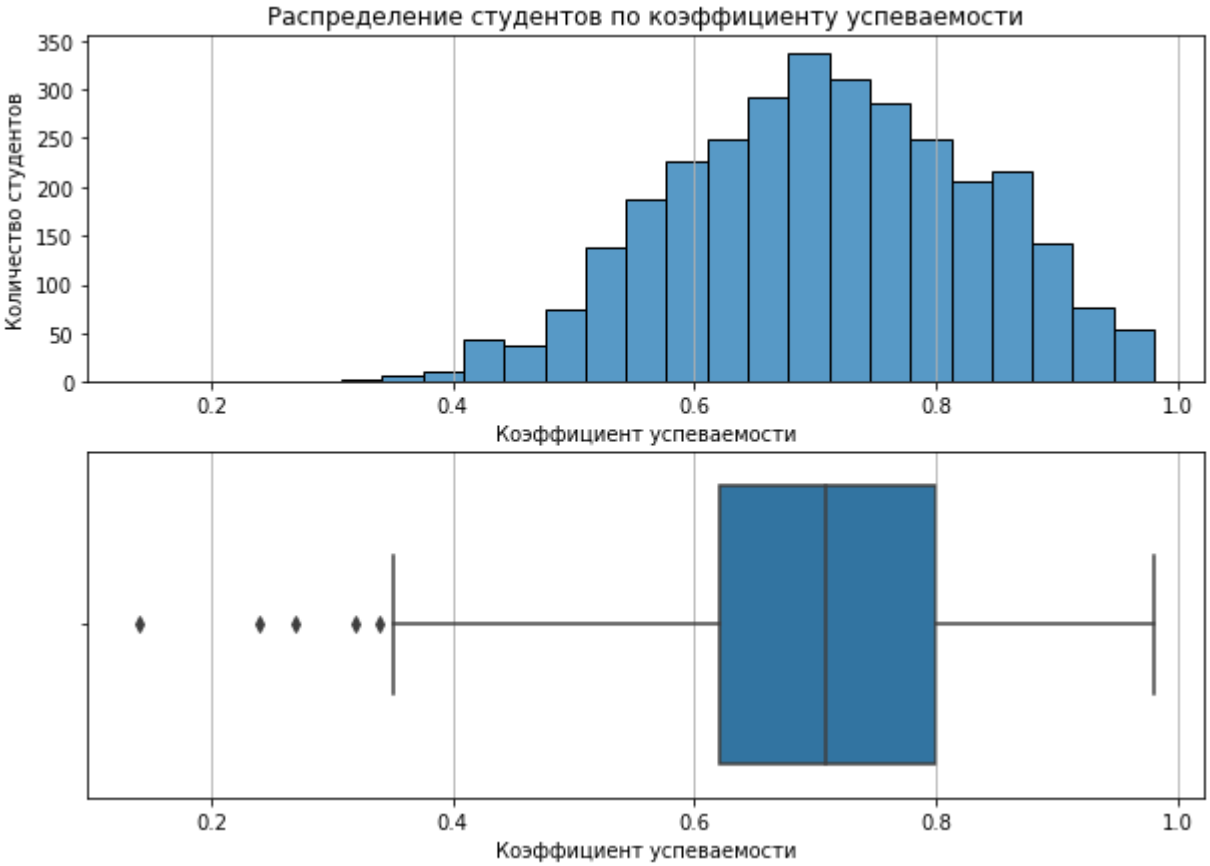
Визуализируем некоторые признаки в другом виде:

Меняли ли студенты когорту



Распределение студентов по сегментам





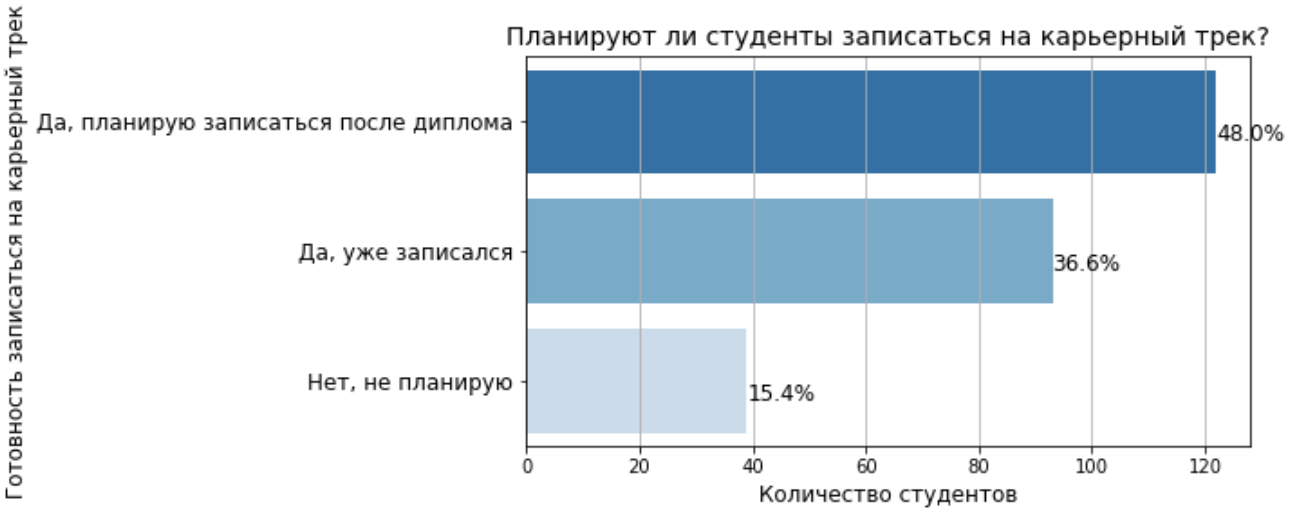
Что можем сказать сразу:

- Большая часть студентов, ответивших на соответствующий вопрос, готовы рекомендовать курс.
- Самые объемные сегменты b2g и b2c. B2b практически не представлен, возможно работодатель просит не включать «Трудоустройство» в курс своим работникам.
- Больше всего студентов учатся на направлениях data analyst и data scientist.
- Большая часть студентов умеренно уверены в полученных ими знаниях, также четверть студентов не могут оценить свои знания.
- Чаще всего студенты доучиваются в тех когортах, в которых начали обучение.
- Случаи перехода из b2g в b2c ещё более редки.
- У половины студентов коэффициент успеваемости от 0.6 до 0.8. Есть отдельные выбросы по этому показателю в диапазоне до 0.4

5.3 Ответы студентов на вопросы:

Посмотрим, как студенты отвечали на вопросы.

При этом в вопросе про бэкграунд объединим варианты 'Нет опыта работы аналитиком и в IT.' и 'Нет опыта работы в IT и в направлении Анализа данных.'



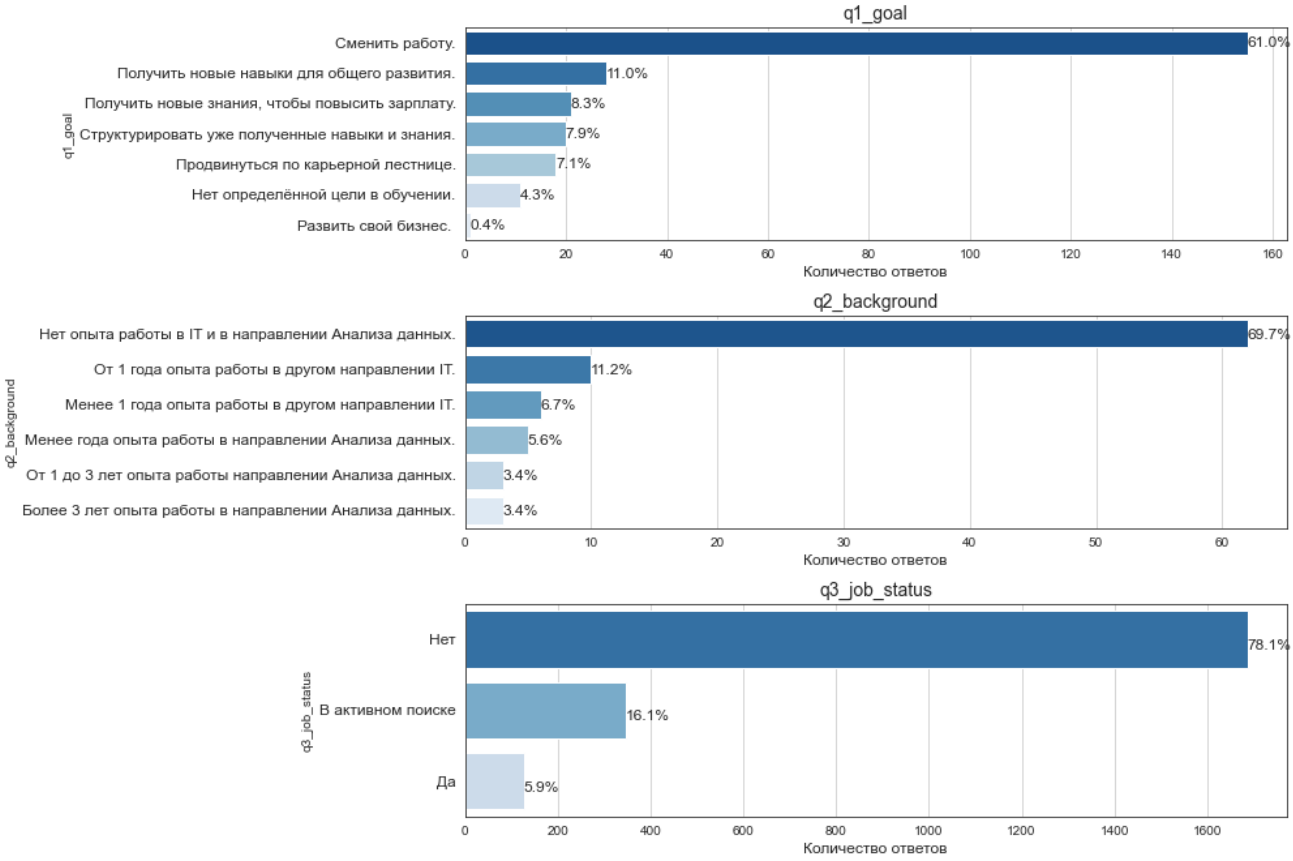
Посмотрим, зависят ли цели студентов от желания записаться на карьерный трек:

Не планируют записываться на карьерный трек студенты, у который основная цель обучения - получить новые навыки для общего развития (31%), продвинуться по карьерной лестнице(18%) или у кого нет определённой цели в обучении (18%). Тем не менее, более 15% таких студентов преследуют цель сменить работу. Попробуем понять, нашли ли уже такие студенты работу:

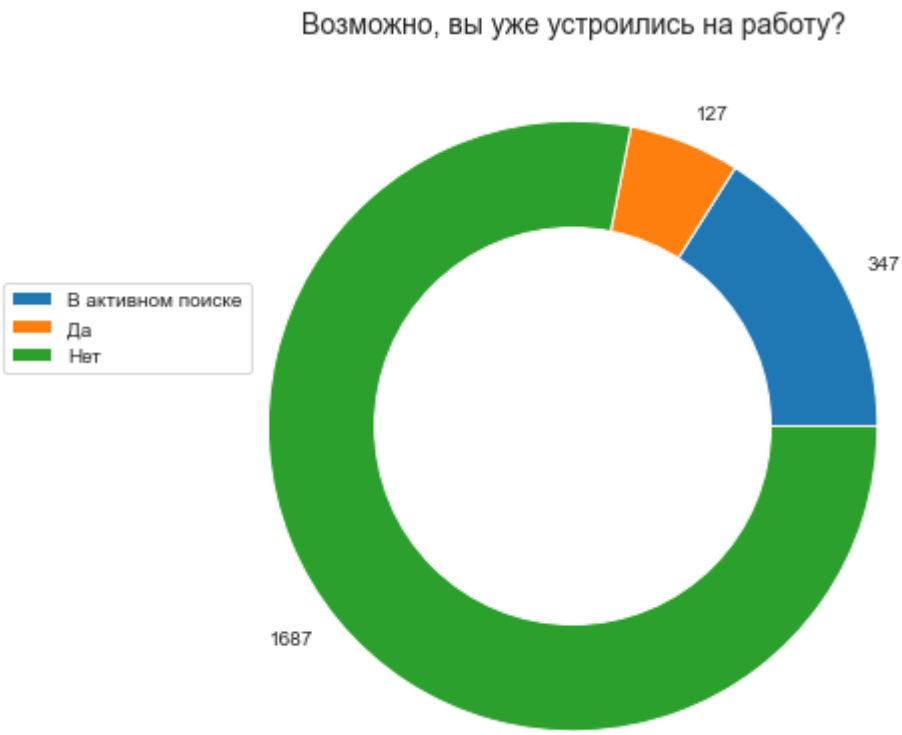
	user_id
q3_job_status	

Студенты, которые ответили на вопрос о желании записаться на карьерный трек отрицательно, не отвечали на вопрос о том, устроились ли они уже на работу во время обучения. Возможно, стоит собрать больше данных о таких пользователях, например задавать им в опроснике дополнительный вопос о причинах отсутствия желания записаться на карьерный трек.

Продолжим обзор ответов на вопросы.



Визуализируем некоторые графики в другом виде:

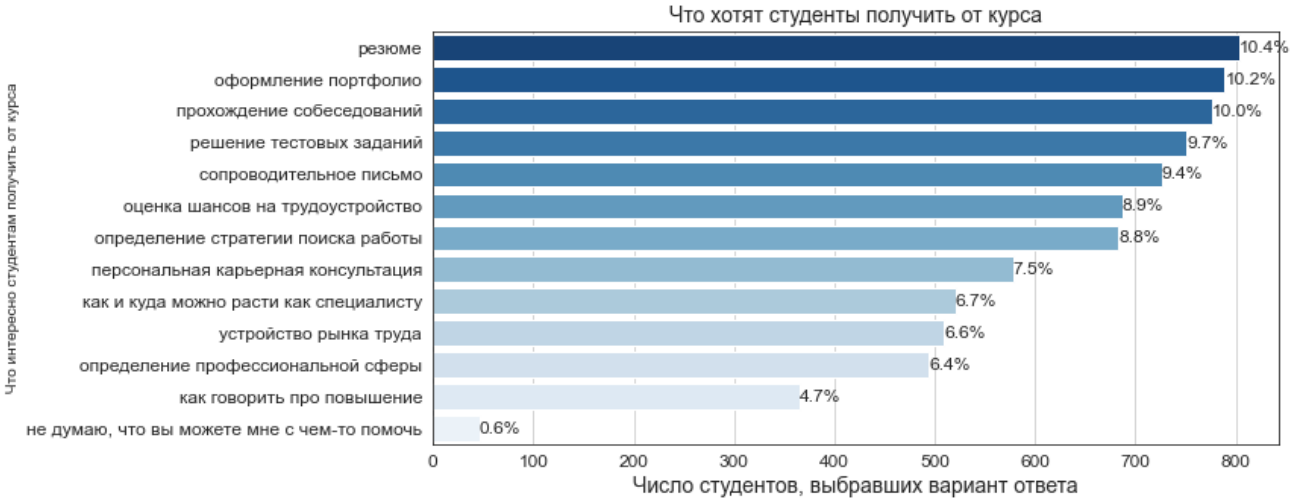


Выводы из ответов на данные вопросы:

1. Большая часть студентов записывалась на курс, чтобы сменить работу. И нет ни одного студента, который бы записывался на курс, чтобы развить свой бизнес.
2. Большая часть студентов записывается на курсы, не имея опыта работы в IT или с минимальным опытом в другом направлении IT.
3. Большая часть студентов на курсе "Трудоустройство" ещё не устроились на работу.
4. На карьерный трек большая часть студентов курса или уже записались, или планируют записаться по окончании диплома (последний вариант популярнее).

5.4 Запросы студентов

Отдельно рассмотрим, как отвечали студенты на вопрос о том, что они хотят получить от курса. Здесь студенты могли выбрать несколько вариантов ответа.



Видим, что в целом студентов больше всего интересуют практические навыки прохождения этапов трудоустройства, такие как:

- составлению резюме
- портфолио
- прохождению собеседований
- решение тестовых заданий
- написание сопроводительных писем

Меньше интересуют стратегические вопросы:

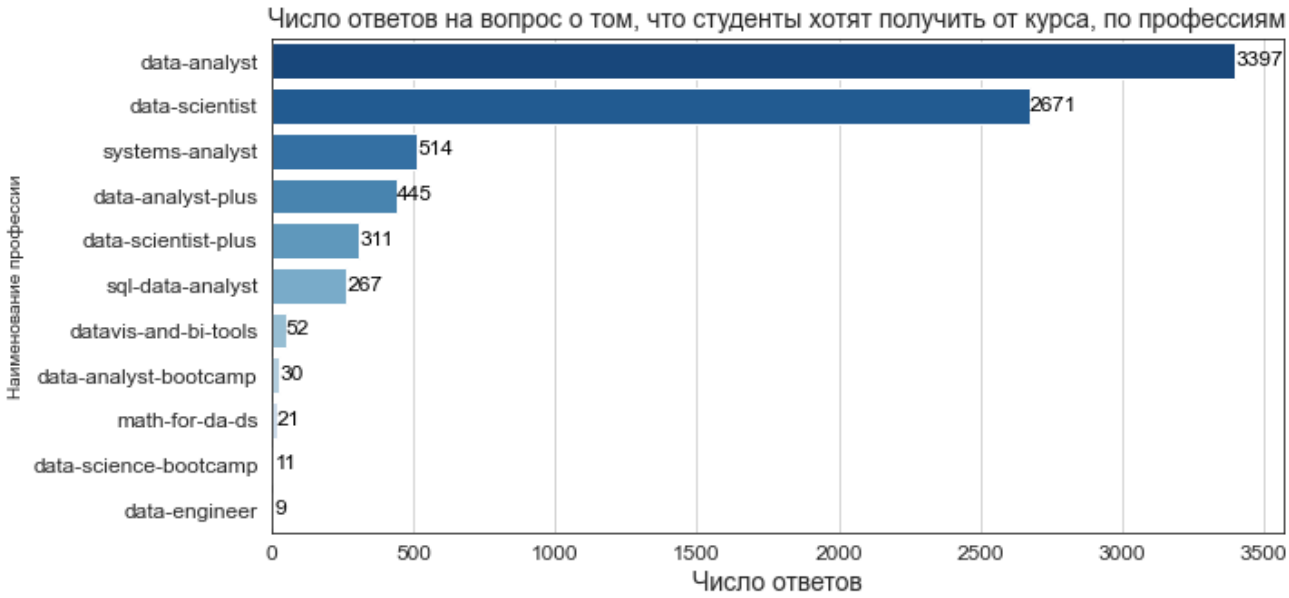
- оценка шансов на трудоустройство
- определение стратегии поиска работы
- персональная карьерная консультация

Меньше всего студентов интересуют более общие вопросы:

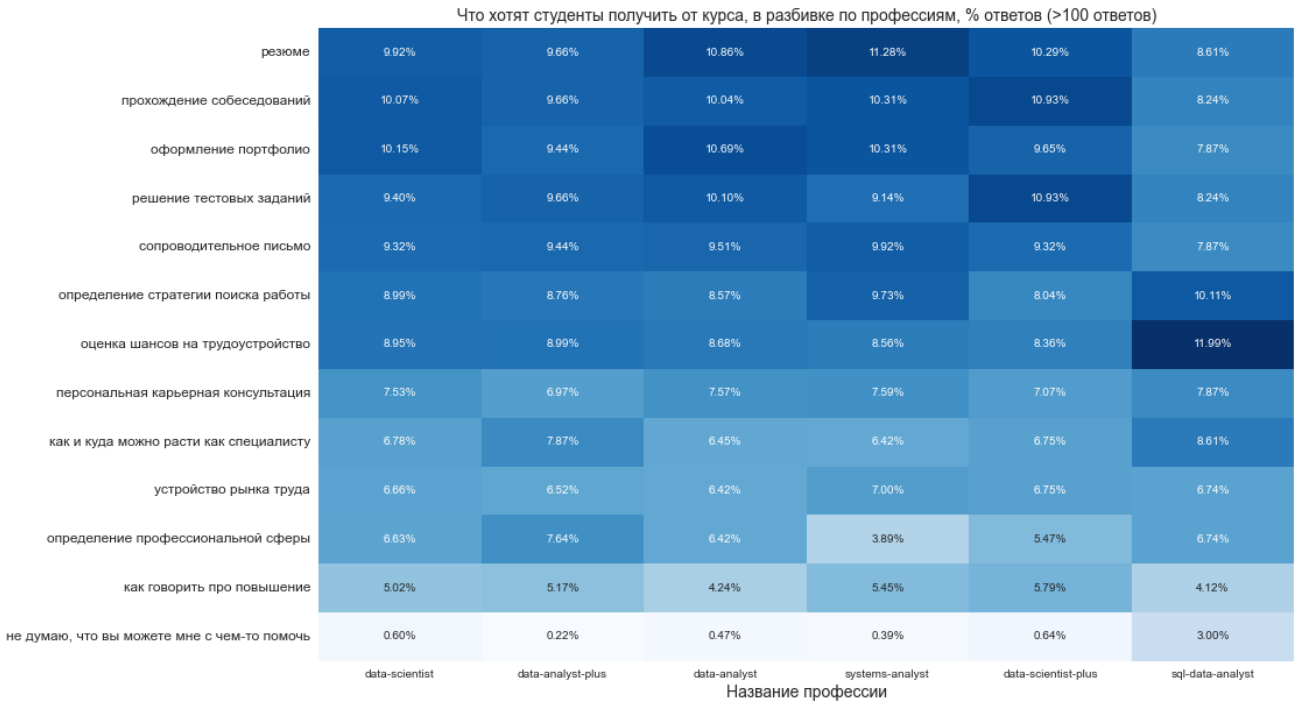
- как и куда расти как специалисту
- устройство рынка труда
- определение профессиональной сферы
- как говорить про повышение

Радует, что студенты настроены позитивно: вариант "не думаю, что вы можете мне с чем-то помочь" выбран наименьшим числом студентов.

Определим число ответов на этот вопрос по профессиям.



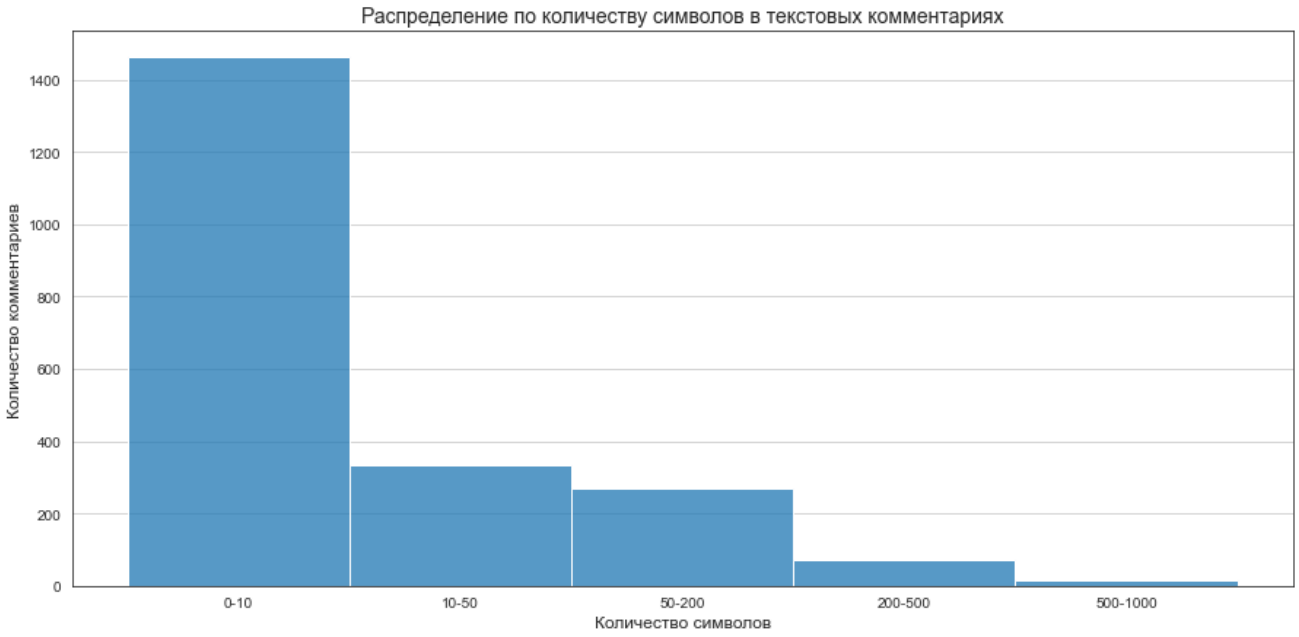
Ответ на этот вопрос будем считать релевантным для профессий, где число ответов не менее 100. Построим график ответов на этот вопрос:



Ответы на вопрос распределены по всем профессиям похожим образом, кроме профессии `sql-data-analyst`: студенты этого курса больше заинтересованы в "стратегических" вопросах: "оценка шансов на трудоустройство" и "опредление стратегии поиска работы". Процент ответа "не думаю, что вы можете мне с чем-то помочь", тоже самый высокий. Видимо, выпускники этого курса чувствуют себя менее уверенно в вопросах трудоустройства.

5.5 Текстовые ответы

Также отдельно рассмотрим текстовые комментарии. Определим, насколько подробные студенты готовы оставлять комментарии.



Большая часть комментариев - до 200 символов. Посмотрим на наиболее часто встречающиеся в них слова, используя wordcloud:

Наиболее часто в комментариях может содержаться информация о том, нашёл студент работу или ещё ищет. Поскольку такой вопрос уже содержался в опроснике, опираться в анализе на эти текстовые комментарии не будем. Но, возможно, в более длинных комментариях содержится более детальная информация о тех трудностях, с которыми сталкиваются студенты.

Посмотрим на комментарии от 200 символов, отберём 5 случайных:

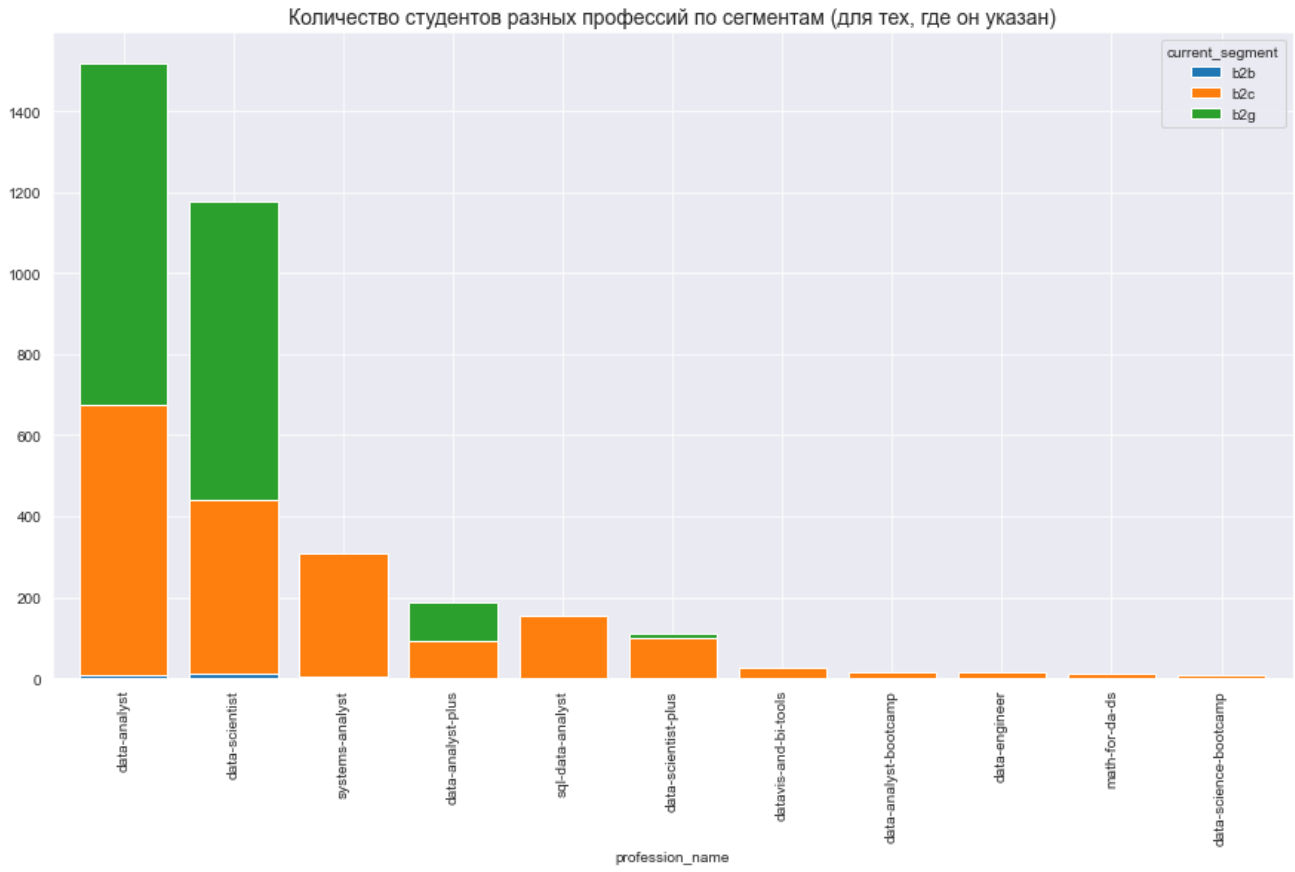
```
['Хотелось бы сменить работу, но пока не менять область профессиональной деятельно
сти кардинально. Возможно, есть какие-то направления на стыке, где мои уже имеющие
ся знания и опыт пригодились бы для лучшей работы.',
 'За время обучения я получила два предложения о работе, но они не соответствовали
моим ожиданиям и не были более интересными, чем моя нынешняя работа по нескольким
параметрам: уровень заработной платы, более удобное расположение относительно мест
а проживания,  наличие социальных гарантий.',
 'У меня был неплохой бэкграунд до начала обучения на ЯП в виде знания MS Excel и
VBA. Ещё до курсов меня взяли на работу статистиком в Департамент Здравоохранения
Москву. А во время обучения повысили до аналитика, и я стал применять Питон вместо
VBA. Мне есть куда расти в профессиональном смысле в Департаменте, но когда я пойм
у, что достиг потолка, мне пригодятся знания и артефакты  с Карьерного трека.',
 'Я работаю преподавателем в вузе, врачом, защитила кандидатскую диссертацию по св
оей специальности. При наличии желания сменить работу (уйти из практической медици
ны) не хотелось бы совсем перечеркивать все, чего я достигла. Может существует как
ое - то направление, где я смогла бы применить не только знания по анализу данных,
но и тот багаж, который у меня уже есть. ',
```

```
 'Я хочу открыть свой бизнес. Знания в области DS нужны только для этого. Мне не п
омешала бы консультация с практикующим специалистом, чтобы уточнить пределы примен
ения в моей сфере деятельности и что ещё изучить, чего не было в курсе. Не уверен
а, что в этом мне поможет данный блок)']
```

В основном люди пишут о себе и о своём опыте, расширяя ответы, данные ранее в опроснике. Возможно, стоит провести отдельное исследование многосимвольных комментариев, которое позволит лучше понять боли пользователей, скорректировать работу коучей, дополнить опросник.

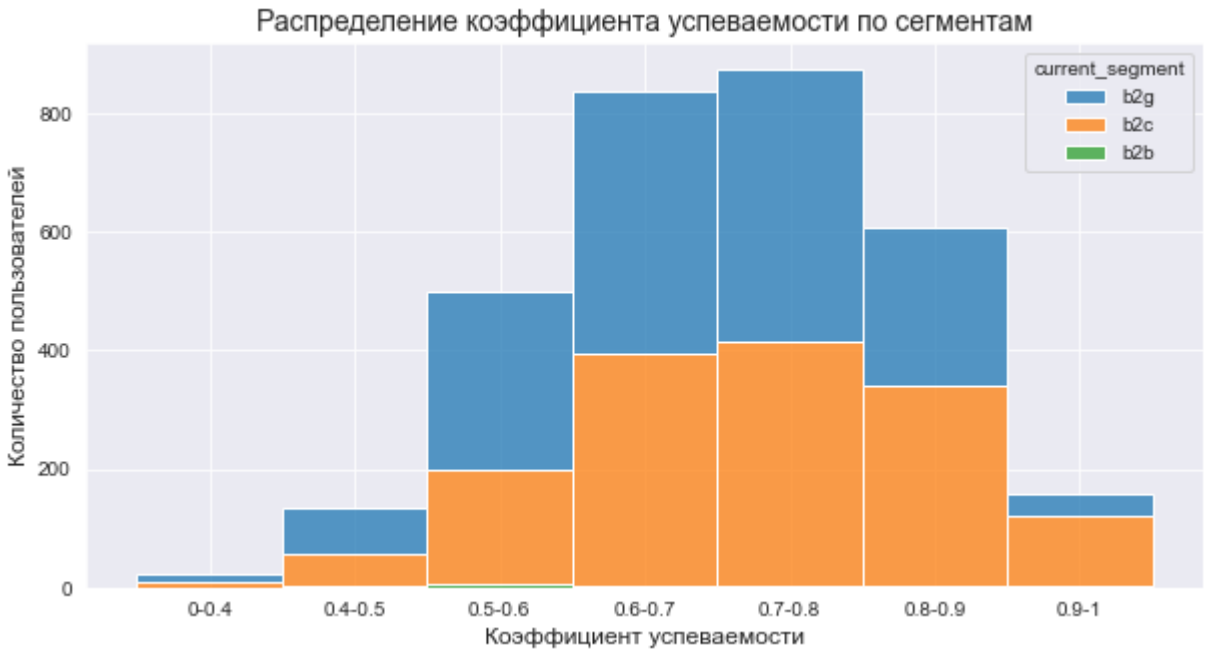
5.6 Особенности сегментов

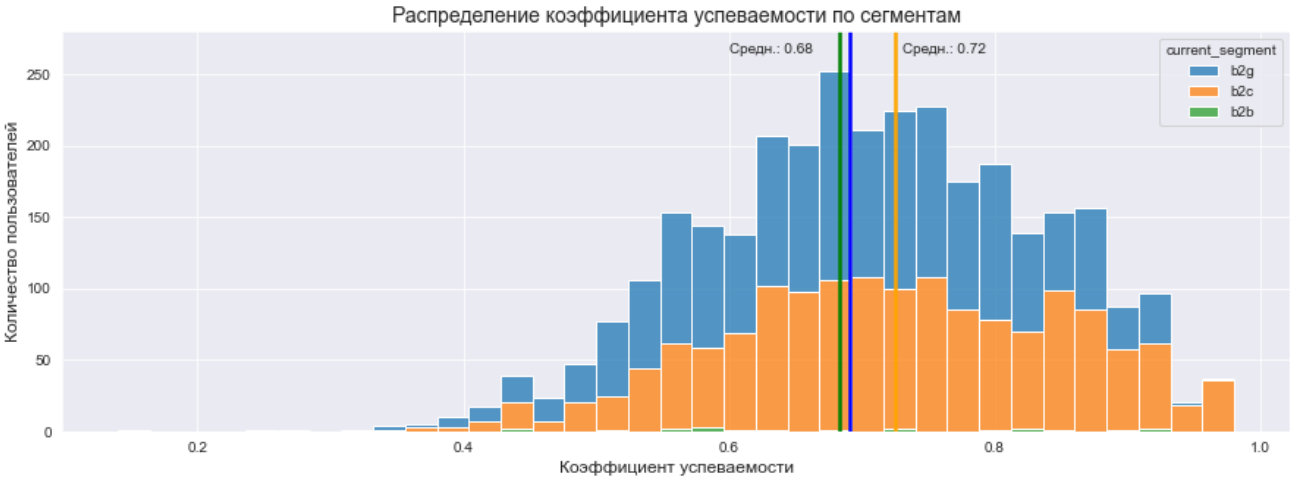
Посмотрим, как пользователи B2B, B2G и B2C распределены по профессиям.



Большинство пользователей сегмента B2G записывались в профессии data-analyst и data-scientist, небольшая доля - в data-analyst-plus и data-scientist-plus. Видимо, остальные курсы не вошли в программу гос.софинансирования. Сегмент b2b представлен крайне слабо, они присутствуют только в профессиях data-analyst и data-scientist. Не будем, впрочем, забывать, что у нас результаты опроса студентов "Трудоустройства", а поскольку b2b студентам не требуется трудоустройство, работодатель иожет иметь возможность блокировать прохождение этого курса своим сотрудникам.

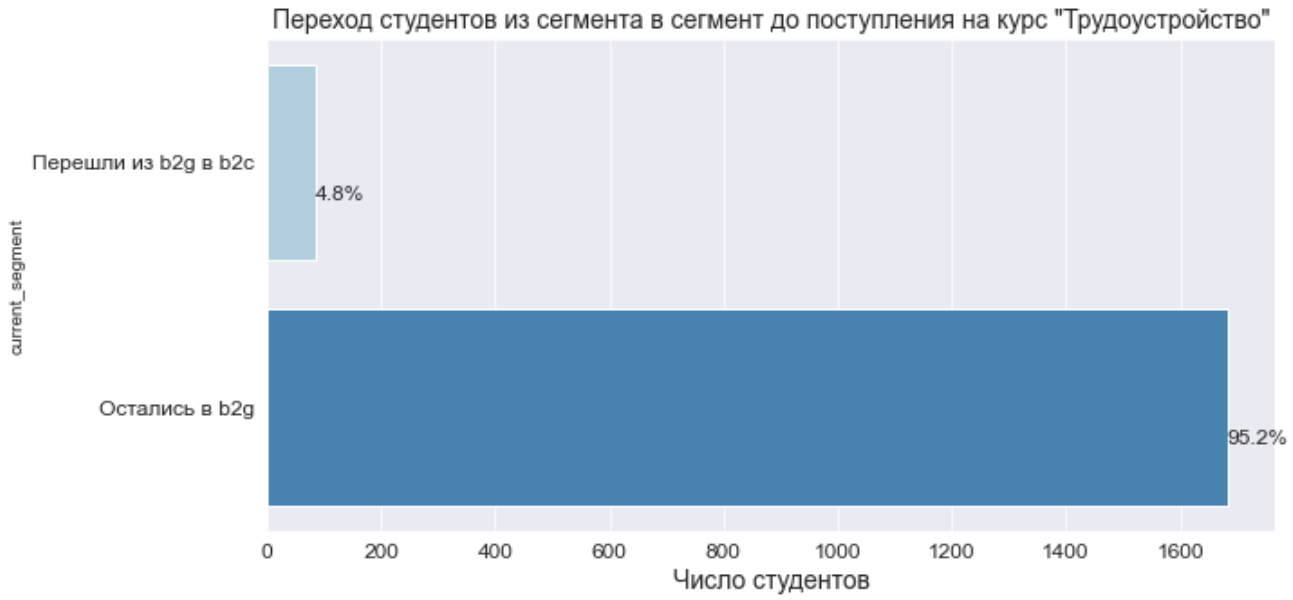
Посмотрим на распределение успеваемости по сегментам:



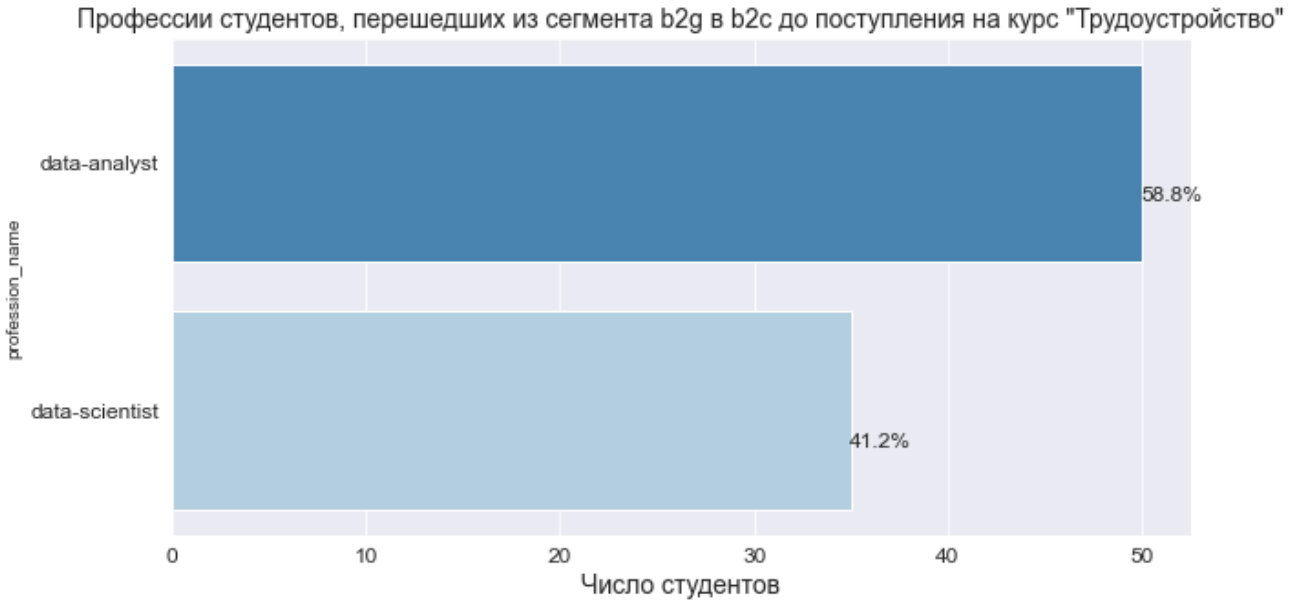


Средняя успеваемость студентов сегмента b2g и b2b немного ниже успеваемости сегмента b2c, что может объясняться дополнительной мотивацией за счёт оплаченного лично курса.

Посмотрим, сколько студентов перешли из b2g в b2c.



Большая часть студентов (95.2%) остались учиться до поступления на курс "Трудоустройство" в b2g, небольшая часть (4.8%) перешли в b2c. Такой переход предположительно связан с невыполненим студентами условий государственой программы, в связи с чем они были вынуждены опатить курс самостоятельно и подолжать его уже в сегменте b2c. Таких студентов немного. Посмотрим, в каких они профессиях:

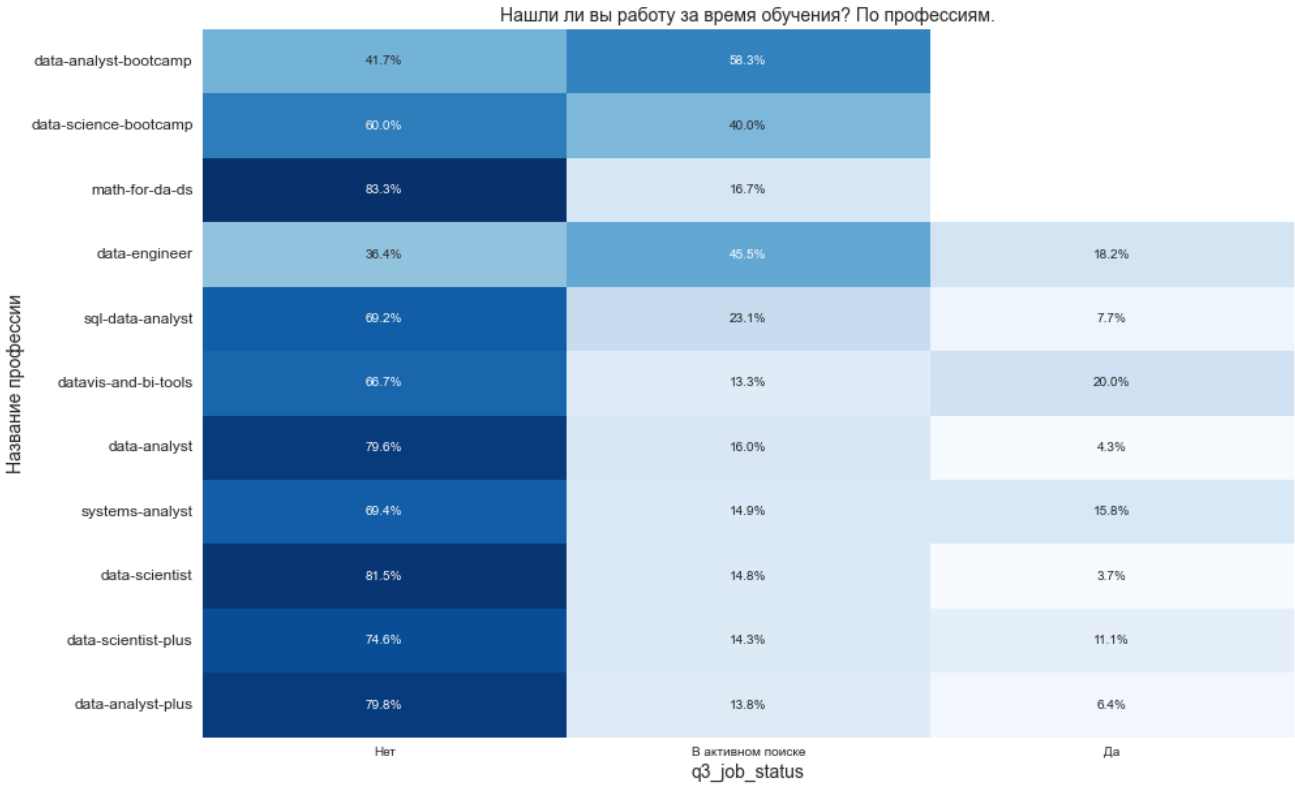


Большая часть студентов, поменявших сегмент, участся профессии `data-analyst`, немногим меньше - `data-scientist`.

5.7 Показатели в разрезе профессий

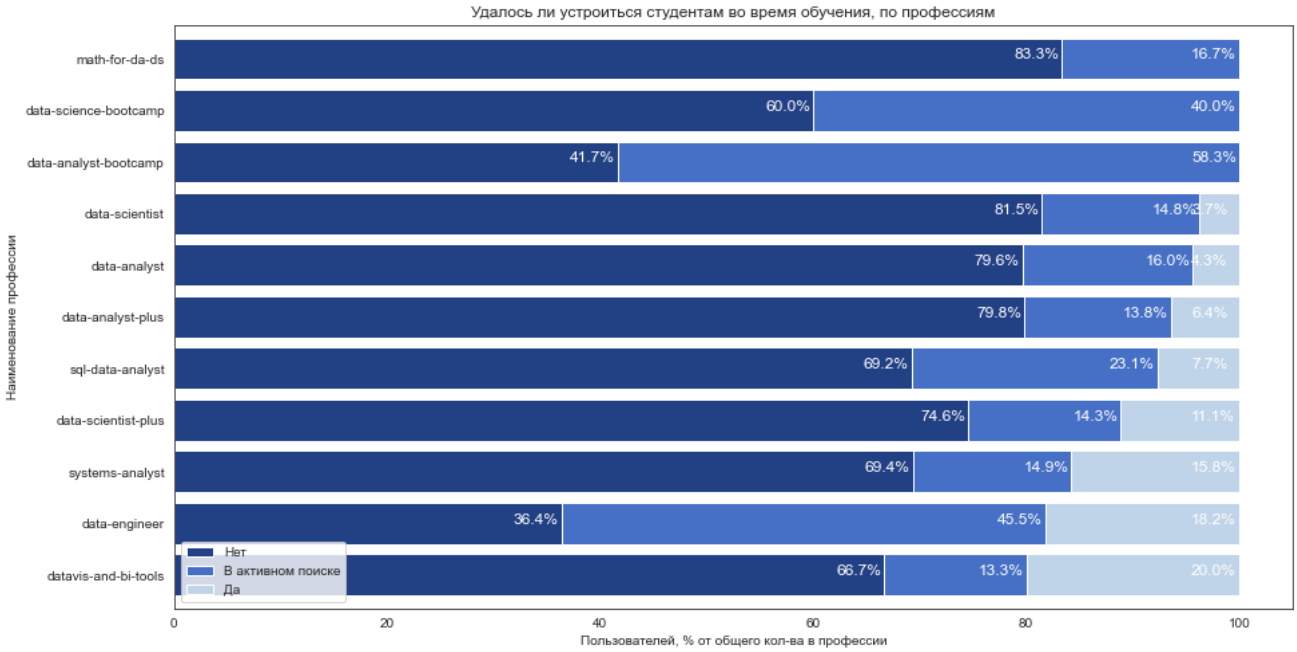
Рассмотрим некоторые показатели в разрезе профессий.

Начнём с вопроса, удалось ли студентам найти работу во время обучения.



Выведем ту же информацию в другом виде.

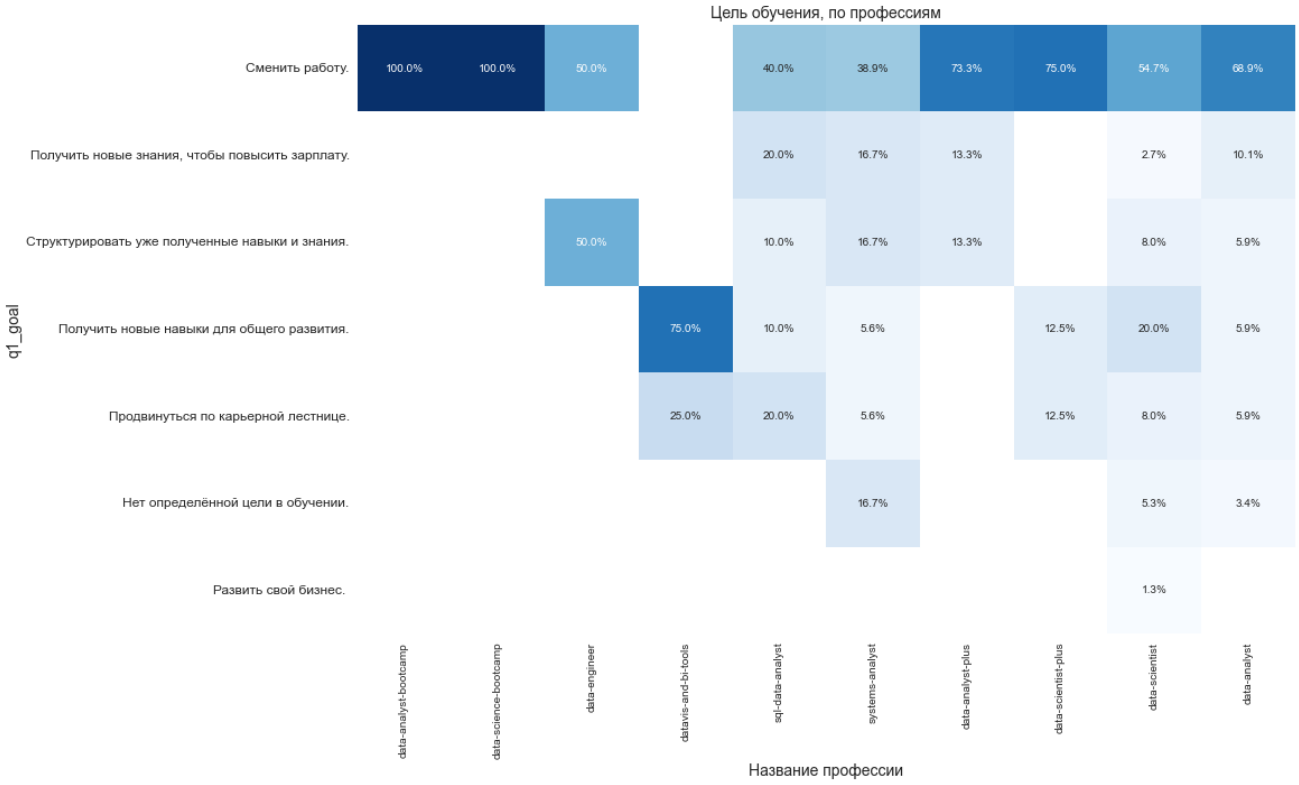
```
<Figure size 720x360 with 0 Axes>
```



Трудоустроенных за время обучения больше всего на курсах **systems-analyst**, **datavis-and-bi-tools**, **data-engineer**, но не более 20%.

Самые активные - студенты буткемпов и курса **data-engineer**

Больше всего не трудоустроенных и при этом не находящихся в активном поиске - **math-for-da-ds**, **data-scietist**, **data-analyst** и **data-analyst-plus**



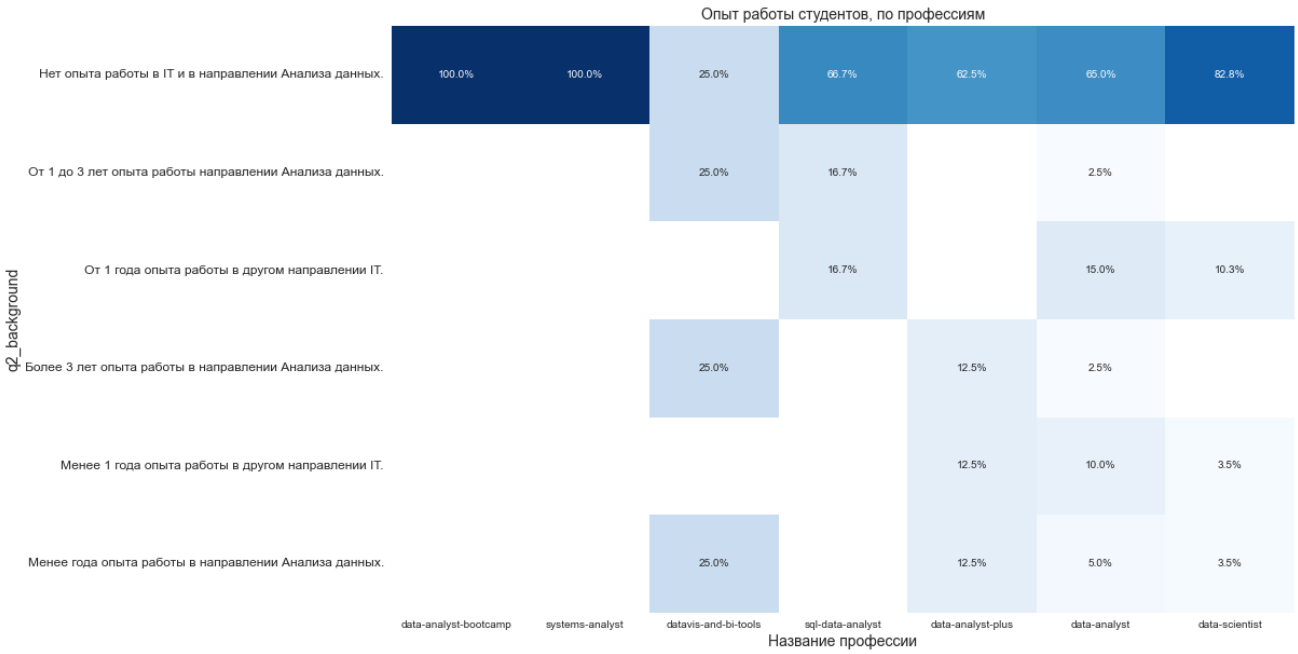
Студенты курса по визуализации данных в первую очередь преследуют цель получить новые навыки для общего развития и продвинуться по карьерной лестнице. Для всех остальных первая цель - сменить работу.

Выраженные второстепенные цели:

- для курса по SQL - продвинуться по карьерной лестнице и получить новые знания, чтобы повысить зарплату.
- для курса data-engineer - структурировать уже полученные навыки и знания
- для курса data-science - получить навыки для общего развития

Для остальных курсов нет выраженной втростепенной цели.

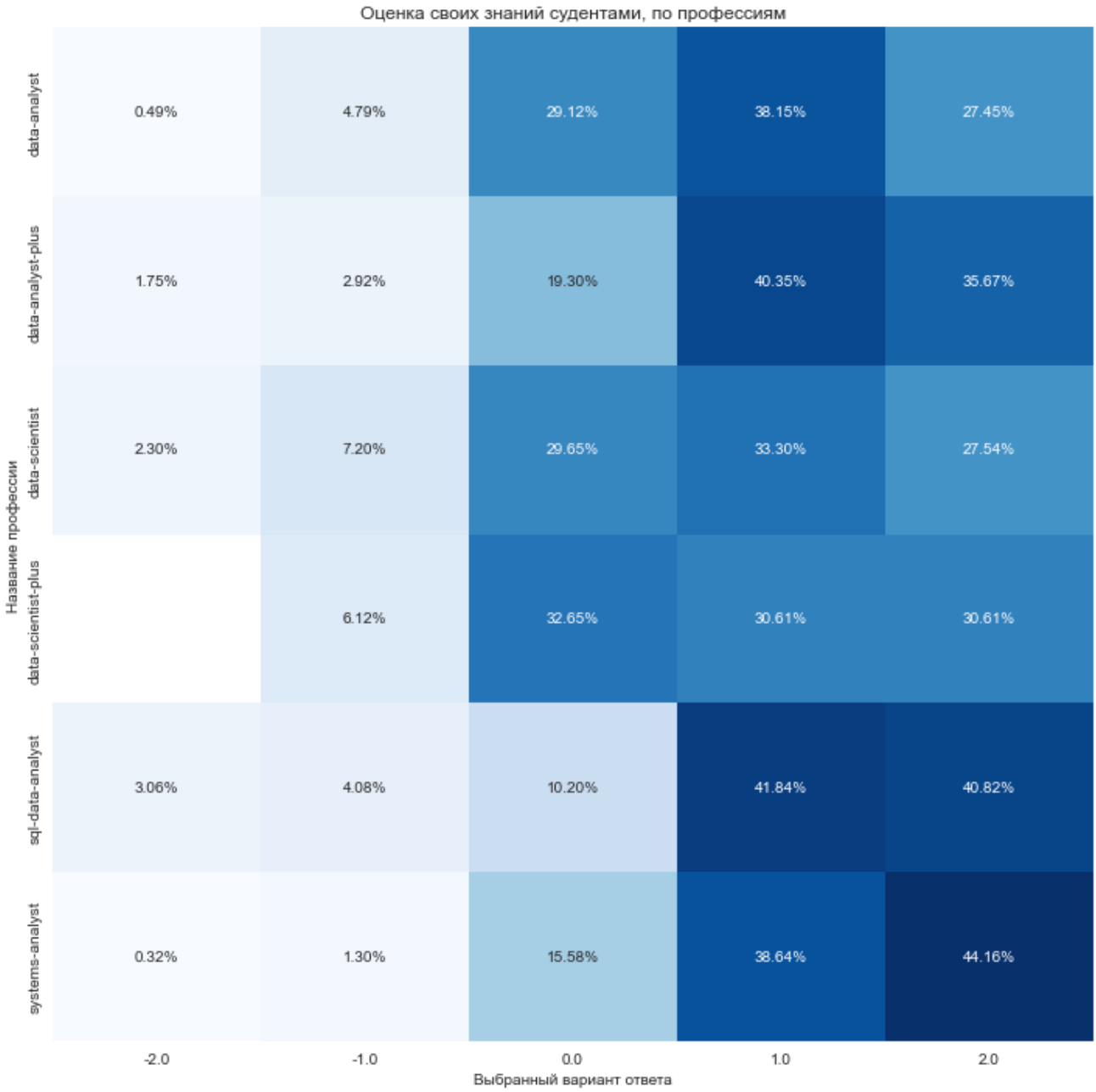
Рассмотрим опыт работы студентов в разрезе профессий. Впрочем, не будем забывать, что крайне мало студентов ответили на этот вопрос.



Для всех профессий наибольшая доля студенов не имеют опыта работы в IT и в направлении Анализа данных.

Исключение - системные аналитики, курс по SQL и курс по визуализации, где больше людей с опытом.

Рассмотрим, как студенты разных профессий оценивают свои знания:

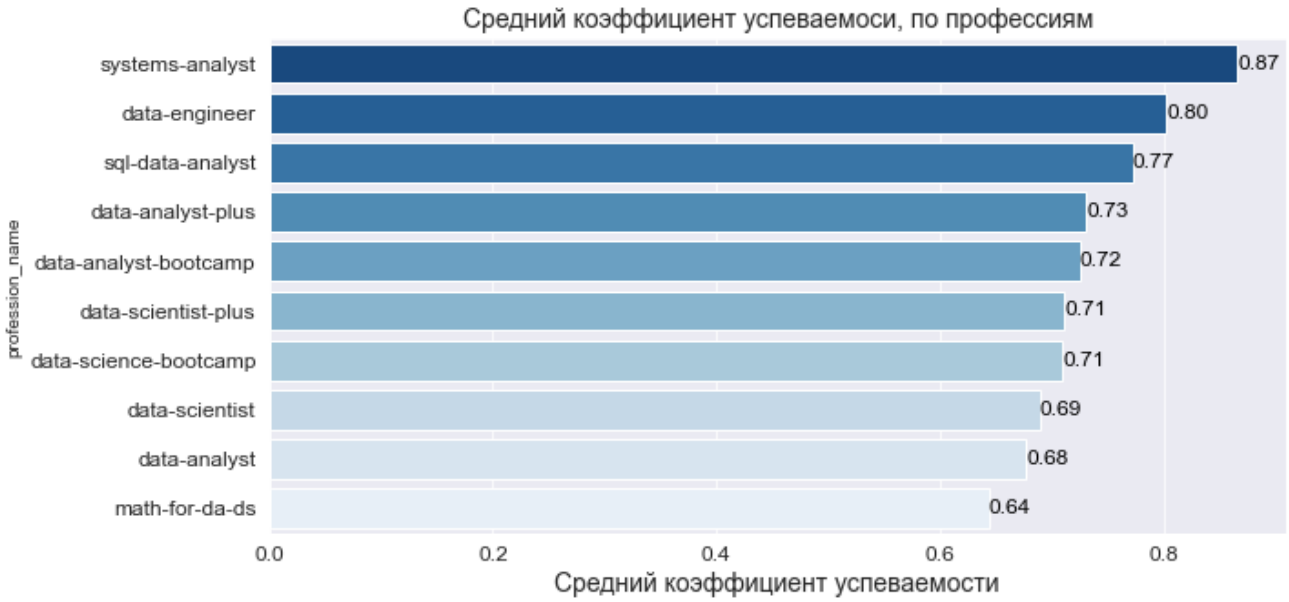


Лучше всего прогрессс в своих знаниях оценивают студенты курсов **systems-analyst, sql-data-analyst** - скорее всего, курсы легки для усвоения навыков.

Вторые по уверенности - **data-analyst, data-analyst-plus**

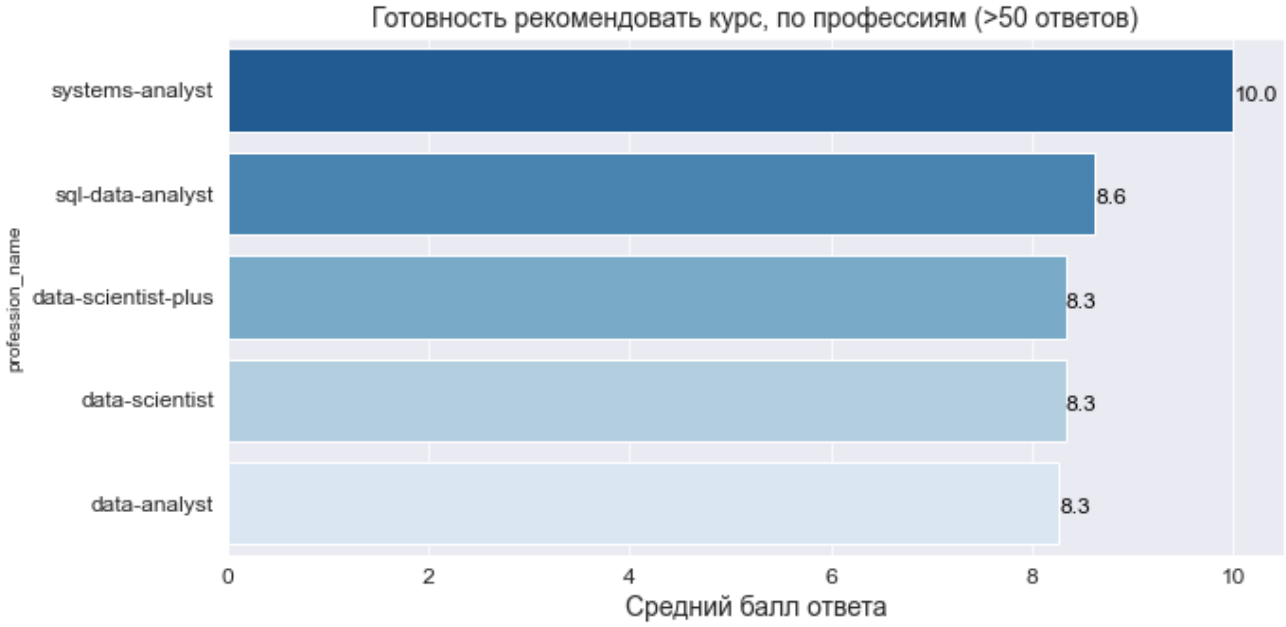
Студенты курсов **data-science и data-science-plus** не так уверены в своём прогрессе в знаниях, видимо материал самый сложный среди аналитических профессий.

Сразу же посмотрим и на объективную оценку успеваемости - на коэффициент, выставленный студентам Практикумом.



Объективно самые успевающие студенты - в профессиях **systems_analyst**, **data_engineer** и **sql_data_analyst**. Впрочем, от них не сильно отстают остальные профессии. Самые сложные для усвоения профессии: **math-for-da-ds**, **data-analyst**, **data-scientist**. Нет данных по курсу визуализации.

Посмотрим, готовы ли студенты рекомендовать курс. Здесь выведем только профессии, где количество оставивших фидбек не ниже 50, во избежание влияния фактора субъективности:

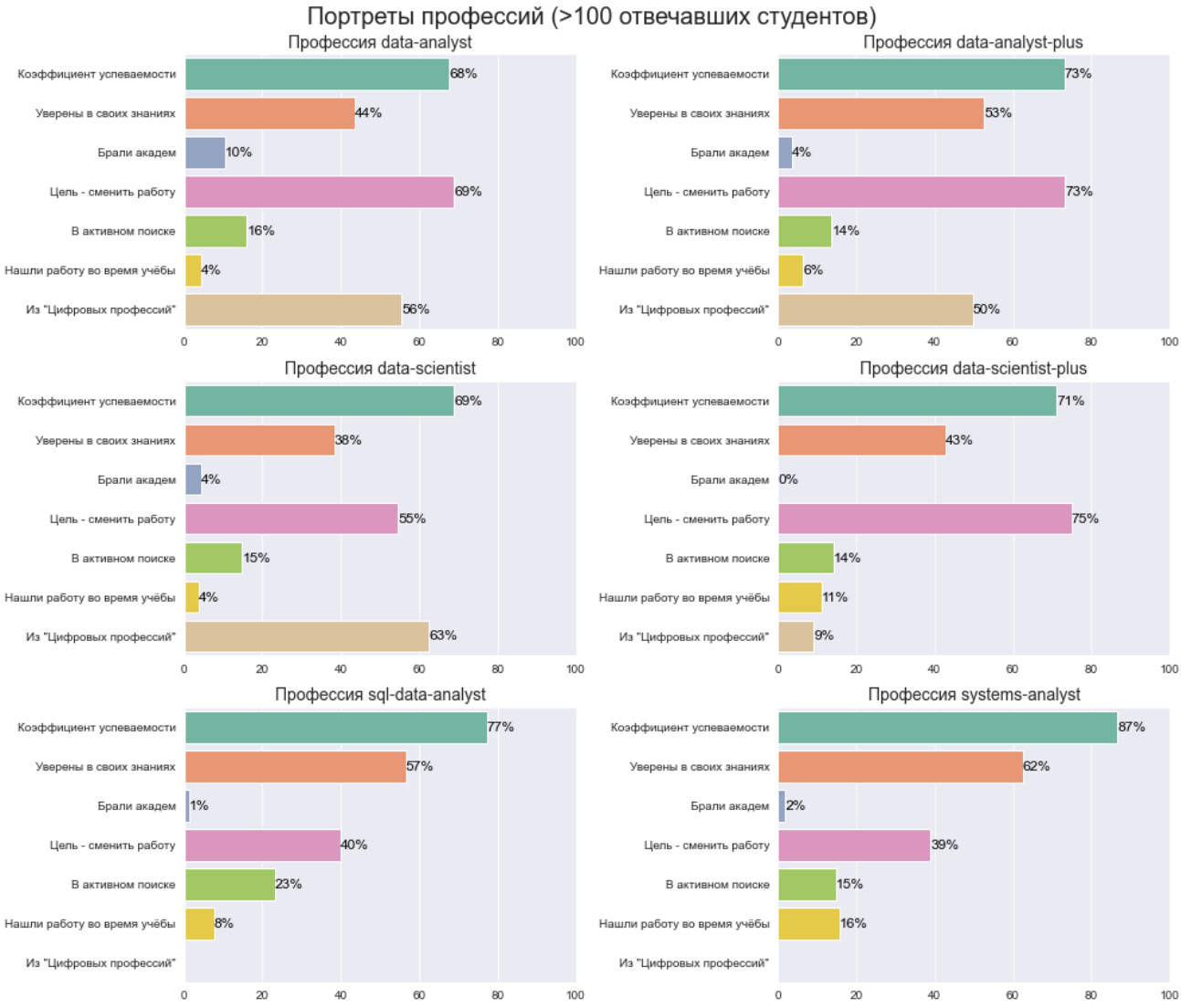


Самые довольные студенты - с курса **systems-analyst**, у остальных средний балл сопоставим.

5.8 Портреты основных профессий

Соберём вместе основные характеристики профессиий с числом отвечавших на опросник студентов >100.

Визуализируем основные характеристики каждой профессии отдельно.



Выводы после сравнения между собой самых популярных профессий:

Специалист по Data Science:

- Самый высокий процент студентов с гос.софинансированием (63%),
- Из профессий с гос.софинансированием самый низкий процент тех, кто песледует главную цель сменить работу (55%).
- Как и в профессии Аналитик Данных, самый низкий процент нашедших работу во время учёбы (4%). Если сравнить с программой "Плюс", в последней он почти в три раза больше (11%).
- Самый низкий процент уверенных в своих знаниях студентов (38%), можно сравнить с программой "Плюс", где уверенных 43%. Один из самых низких коэффициентов успеваемости 68% (наряду с профессией "Аналитик данных") это подтверждает.

Программа, скорее всего, сложная: студенты не чувствуют уверенности в своих знаниях, им трудно устраиваться на работу.

Специалист по Data Science Плюс: По сравнению с обычной программой:

- Самый малый процент студентов с гос.софинансированием из тех профессий, где она была (9%)
- На два процентных пункта выше коэффициент успеваемости (71%)

- Уверены в своих знаниях на 5 процентных пункта больше студентов (43% студентов, но тоже процент не высокий)
- Нет бравших академ студентов
- Сменить работу преследует цель 75% студентов - самый высокий процент среди всех основных профессий
- В три раза больше процент нашедших работу во время обучения (11%)

Видимо, студенты чувствуют себя несколько увереннее, чем в обычной программе, и гораздо быстрее находят работу.

Аналитик данных:

- Самый низкий коэффициент успеваемости (68%), сравним с профессией "Специалист по Data Science* (69%), но при этом уверенность студентов в своих знаниях выше - 44% против 38%
- Самый высокий процент бравших академ студентов - каждый десятый.
- Один из самых высоких процентов по цели обучения "смена работы" - 69% студентов
- Наряду с профессией "Специалист по Data Science", самый низкий процент нашедших работу во время обучения (4%)

Вопреки распространённому мнению, наши данные говорят о том, что профессия не проще в усвоении, чем "Специалист по Data Science", здесь даже больше студентов, которые не успевают её осваивать, находят работу студенты так же тяжело. Тем не менее, чувствуют себя увереннее в собственных знаниях, чем в Data Science - возможно, это связано с отсутствием столь высоких требований к знаниям математики.

Аналитик данных плюс: По сравнению с обычной программой

- Коэффициент успеваемости выше на 5 процентных пунктов. Студенты реже берут акадм (4% против 10% в обычной программе)
- На 9% выше процент уверенных в своих знаниях (53%, самый высокий показатель среди всех профессий, которые участвовали в программе гос.софинансирования)
- На 2 процентных пункта больше студентов, которые нашли работу во время учёбы. Процент всё же невысокий.

Студенты здесь увереннее, чем в обычной программе, лучше усваивают программу, реже берут академы, но на работу устраиваются лишь немногим легче. Если сравнивать с Data Science, где с ростом уверенности в знаниях растёт и возможность трудоустройства, возможно, здесь сложнее трудоустроиться скорее из-за высокой конкуренции: аналитиков данных среди выпускников аналитических профессий - самое большое количество.

SQL для анализа данных:

- Осваивают программу студенты очень хорошо (77%) и почти не брали академы (1%)
- Студенты весьма уверены в своих знаниях (57%)
- Цель сменить работу преследуют только 40% студентов - один из самых низких показателей.
- Тем не менее, в активном поиске её 23% и 8% нашли работу во время учёбы.

Программа достаточно проста в усвоении, скорее всего её проходят либо чтобы подтянуть знания, либо как первую ступень на пути к анализу данных - отсюда возможно и идёт самый маленький процент желающих сменить работу, но в то же время больше всего других профессий тех, кто в активом поиске или нашёл работу во время учёбы.

Профессия Системный аналитик:

- Самый высокий коэффициент успеваемости - 87%, студенты почти не берут академы (2%)
- Уверены в своих знаниях 62% студентов, самый высокий показатель
- Цель сменить работу преследуют 39% студентов, самый низкий показатель
- В то же время в активном поиске 15% студентов (обычный показатель), а нашедших работу во время учёбы - 16% студентов (самый высокий среди основных профессий)

Материал студенты усваивают легче, чем в других профессиях, и они уверенне других в своих знаниях. К сожалению, очень мало студентов ответили на вопрос об опыте работы, но мы можем предположить, что сюда идёт больше людей с опытом, отсюда меньше людей, пришедших за сменой работы. В то же время, находят работу здесь студенты легче, чем остальных основных профессиях. Самая позитивная история.

6. Общие выводы

Выводы предобработки

- Студенты часто пропускают вопросы, не отвечая на них.
- Студенты отвечают на вопросы регуляно, в этом случае мы в анализе учитываем последний ответ.
- Один студент может одновременно обучаться нескольким профессиям, в этом случае мы учитываем последние ответы студента по каждой профессии.

Выводы исследовательского анализа

- Больше всего студентов учатся на направлениях data analyst и data scientist.
- Средняя оценка готовности порекомендовать курс увеличивается с ростом уверенности студента в своих знаниях.
- Самые объемные сегменты b2g и b2c. B2b практически не представлен (возможно, b2b не даёт возможности своим сотрудникам поступать на курс "Трудоустройство").
- Большая часть студентов (95.2%) остались учиться до поступления на курс "Трудоустройство" в b2g, небольшая часть (4.8%) перешли в b2c. Такой переход предположительно связан с невыполненим студентами условий государственой программы, в связи с чем они были вынуждены опатить курс самостоятельно и подолжать его уже в сегменте b2c.
- Случаи перехода из b2g в b2c ещё более редки. Большая часть студентов, поменявших сегмент, участся профессии data-analyst, немногим меньше - data-scientist.
- Самые довольные студенты - с курса systems-analyst, у остальных средний балл сопоставим.

Выводы из ответов пользователей на вопросы

- Большая часть студентов записывалась на курс, чтобы сменить работу. И нет ни одного студента, который бы записывался на курс, чтобы развить свой бизнес.
- Большая часть студентов записывается на курсы, не имея опыта работы в IT или с минимальным опытом в другом направлении IT.
- Большая часть студентов на курсе "Трудоустройство" ещё не устроились на работу.
- На карьерный трек большая часть студентов курса или уже записались, или планируют записаться по окончании диплома (последний вариант популярнее).
- Большая часть студентов, ответивших на соответствующий вопрос, готовы рекомендовать курс. Однако число ответивших небольшое и поскольку студенты сами решали, отвечать на него или нет, ответы могут быть нерелевантными
- Большая часть студентов уверены в своих знаниях
- В текстовых комментариях основном люди пишут о том, нашли или не нашли работу (короткие комментарии), либо расширяют ответы, данные ранее в опроснике. Возможно, в дальнейшем стоит провести отдельное исследование многосимвольных комментариев, которое позволит, исходя из ответов пользователей и их содержательных комментариев, вынести выводы об их основных потребностях в курсе, что может быть полезно для будущей коррекции опросника.

Что интересует пользователей

В целом студентов больше всего интересуют практические навыки прохождения этапов трудоустройства, такие как:

- составлению резюме
- портфолио
- прохождению собеседований
- решение тестовых заданий
- написание сопроводительных писем

Меньше интересуют стратегические вопросы:

- оценка шансов на трудоустройство
- определение стратегии поиска работы
- персональная карьерная консультация

Меньше всего студентов интересуют более общие вопросы:

- как и куда расти как специалисту
- устройство рынка труда
- определение профессиональной сферы
- как говорить про повышение

Радует, что студенты настроены позитивно: вариант "не думаю, что вы можете мне с чем-то помочь" выбран наименьшим числом студентов.

Выводы по основным профессиям

- **Специалист по Data Science**: Программа сложная: студенты не чувствуют уверенности в своих знаниях, им трудно устраиваться на работу.

- **Специалист по Data Science Плюс**: По сравнению с обычной программой студенты чувствуют себя несколько увереннее, чем в обычной программе. С ростом компетенций, которыми здесь удаётся овладеть в более полной мере, растёт и процент трудоустроенных за время обучения (12% против 4% в обычной программе)

- **Аналитик данных**: Вопреки распространённому мнению, данные говорят о том, что студентам здесь не проще, чем в Data Science: даже ещё больше студентов, которые не успевают осваивать программу, находят работу студенты так же тяжело (4% трудоустроенных за время обучения). Тем не менее, студенты в собственных знаниях здесь увереннее, чем в Data Science - возможно, это связано с отсутствием столь высоких требований к знаниям математики.

- **Аналитик данных плюс**: Студенты здесь увереннее, чем в обычной программе, лучше усваивают программу, реже берут академы, но на работу устраиваются лишь немногим легче. Если сравнивать с Data Science, где с ростом уверенности в знаниях растёт и возможность трудоустройства, возможно, здесь сложнее трудоустроиться скорее из-за высокой конкуренции: аналитиков данных среди выпускников проанализированных нами профессий - самое большое количество.

- **SQL для анализа данных**: Программа проста в усвоении, скорее всего её проходят чтобы подтянуть знания, отсюда возможно и идёт самый маленький процент желающих сменить работу, но в то же время в сравнении с другими профессиями больше тех, кто в активом поиске или нашёл работу во время учёбы.

- **Системный аналитик**: Материал студенты усваивают легче, чем в других профессиях, и они уверенне других в своих знаниях. К сожалению, очень мало студентов ответили на вопрос об опыте работы, но мы можем предположить, что сюда идёт больше людей с опытом, отсюда меньше людей, пришедших за сменой работы. В то же время, находят работу здесь студенты легче, чем остальных основных профессиях. Самая позитивная история.

Выводы из сегментации пользователей

- Самый *распространенный* сегмент пользователей "Яндекс Практикума" представляют люди без опыта в ИТ и изучаемой профессии, которые хотели бы сменить работу. Подавляющее большинство студентов **DA, DS, DA-plus, DS-plus** являются представителями данного сегмента.
- Выделяются пользователи, которые не планируют менять работу и пользоваться услугами карьерного трека. Их основная задача - структурировать знания и добиться повышения з/п на текущем месте работы. Большое количество пользователей данного типа обучаются по направлению **Systems Analyst**.
- Третий тип пользователей отличается наличием опыта работы по выбранной специальности и тем, что их главной задачей является продвижение по карьере и повышение з/п. В отличие от представителей второго сегмента, студенты из данной категории готовы сменить место работы и зарегистрироваться в карьерном треке. Существенная доля пользователей из данного сегмента обучается по направлениям **Визуализация данных** и **SQL Data Analyst**. В то же время, из ответов на вопросы студентов последнего следует, что они больше заинтересованы в "стратегических"

вопросах: "оценка шансов на трудоустройство" и "опредление стратегии поиска работы". Процент ответа "не думаю, что вы можете мне с чем-то помочь" самый высокий по сравнению с другими профессиями. Видимо, выпускники этого курса чувствуют себя менее уверенно в вопросах трудоустройства.