

# Домашнее задание №1

## Описание задания

Вам предоставлены тестовые данные по клиентам банка, которые имеют кредит.

Задание состоит из 2-ух частей:

1. Исследование данных и обработка данных для проведения последующей сегментации;
2. Составить профили клиентов на основе проведенных сегментаций

(Использовать минимум 2 метода сегментации).

Каждый студент выбирает вариант, который указан напротив его ФИО в списке

[https://docs.google.com/spreadsheets/d/1jIY6dqZmYKZq65e\\_pcbHFOIL31KK6Gd\\_xoiP9IteMY/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1jIY6dqZmYKZq65e_pcbHFOIL31KK6Gd_xoiP9IteMY/edit?usp=sharing)

Варианты и описание данных представлены в папке: [https://drive.google.com/open?id=1jxwwFMEHVh91ZKx5\\_PqnMSUlsIcytlbI](https://drive.google.com/open?id=1jxwwFMEHVh91ZKx5_PqnMSUlsIcytlbI)

**Для того, чтобы получить оценку, требуется**

1. Прислать архив с файлами, где производились все расчеты и сопроводительное письмо с выводами и комментариями по каждой части:
  - Расчеты могут производиться через код (python/sas/sql), сводные таблицы и формулы в excel или проект SAS Viya;
  - Все выводы необходимо подтверждать визуально интерпретируемыми графиками и данными.
2. Архив (.zip) с файлами требуется отправить на почту [Natalia.Titova@sas.com](mailto:Natalia.Titova@sas.com) с темой «ФКН ВШЭ»
3. Название файла требуется отправлять по шаблону <Имя>\_<Фамилия>\_<номер группы>\_hw1.zip.

Пример, Alexander\_Sharipov\_156\_hw1

# Домашнее задание №1

## Описание задания (Часть 1)

По полученным данным необходимо провести исследование данных:

### 1. Исследуем распределения по данным:

- Рассчитываем кол-во уникальных значений, нулевых и пустых значений + доля в % от общего кол-ва;
- Среднее значение, медиана, стандартное отклонение, минимум, максимум, тип данных по каждому показателю в предоставленных данных;
- Исследуем распределение данных по полу, возрасту и другим категориальным показателям;

### 2. Делаем проверку на:

- Полноту данных по клиентам;
- Пропущенные и нулевые значения в полях;
- Наличие некорректных знаков;

### 3. Готовим итоговую витрину данных для сегментации, при необходимости:

- Корректируем данные – исправляем ошибки;
- Исключаем клиентов с большим числом пропусков или восстанавливаем пропущенные значения;
- Переводим категориальные показатели в целочисленные;

### 4. Описываем все пояснения по исследованию данных и по всем преобразованиям данных.

### Результатом выполнения части 1 будет:

- Таблицы, сводные отчеты и графики (Например, гистограммы и диаграммы Бокса и Вискера);
- Визуализации предоставить в файлах Python/Excel/SAS;
- **Все данные должны быть сопровождаемы выводами по полученному результату исследования (3-4 предложения);**
- Финальная витрина для построения сегментации.

# Домашнее задание №1

## Описание задания (Часть 2)

Провести любыми 2-мя методами сегментацию данных на выбор:

- Бизнес-правила;
- Квантили (RFM);
- Сегменты на основе данных кластеризации с учителем (Дерево решений, регрессия, нейросети, градиентный бустинг и тд.);
- Без учителя (Метод К-средних, Mean-shift, DBSCAN, Иерархическая кластеризация и тд.)

**Задачи, которые необходимо решить в части 2:**

- Выделить сегменты клиентов согласно выбранному методу сегментации;
- Сформировать портреты клиентов на основе полученных данных - дать интерпретацию каждому полученному сегменту;
- Обосновать выбор метода - описать плюсы/минусы каждого метода на основе испытываемых данных и на основе полученного теоретического материала;

P.S. Описание выбора метода сегментации «я его выбрал потому, что проще сделать и только его знаю» –не подойдет!

### Помните!

Вне зависимости от выбранного метода сегментации, выделение групп клиентов должно соответствовать следующими условиями:

- Внутри сегмента однородность максимальная;
- Между сегментами однородность минимальна

### Помните!

Если будете выполнять сегментацию при помощи кода (Python) :

- Для преобразования категориальных переменных, которые имеют различные значения/названия, применяется **One-Hot Encoding**.
- Перед построением сегментации **проводите нормализацию и PCA**

# Домашнее задание №1

## Система оценки

Результат за решение измеряется по 10-балльной шкале, где

**«8-10»** — задание решено полностью, выполнены все 2 части домашней работы:

- проведен анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
- построены сегментации *2-мя методами*;
- предоставлены понятные выводы с подтверждёнными данными (таблицы, графики);

**«6-7»** — задание решено не полностью или с недочётами:

- проведен анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
- построена сегментация *хотя бы одним методом*;
- предоставлены понятные выводы с подтверждёнными данными (таблицы, графики);

**«4-5»** — задание решено с существенными недочётами,

- проведен анализ данных, предоставлен рабочий код и таблицы по исследованию данных;
- выявлены верхнеуровневые зависимости и закономерности по клиентам без построения модели сегментации;

**«0-3»** — задание не решено или решено неверно.

### Помните!

Два одинаковых или почти одинаковых присланных кода и вывода будут оштрафованы:  
**0 баллов за задание независимо от результатов.**

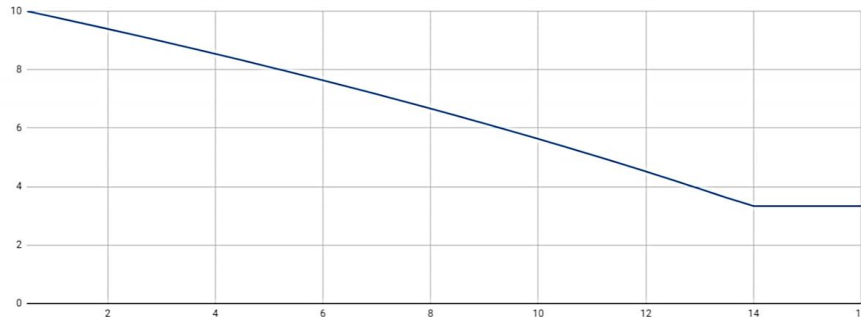
### Система оценки


Решения, присланные после даты дедлайна, будут оштрафованы: максимальный балл за ДЗ будет снижаться в зависимости от величины просрочки дедлайна (в днях):

$$\text{МаксимальныйБалл} = \max\left(\frac{1}{3}, \log_2\left(1.23 + 0.77 * \max\left(0, \left(1 - \frac{\text{ДнейПослеДедлайна}}{14}\right)\right)\right)\right).$$

График зависимости максимально возможного балла за ДЗ от просрочки дедлайна приведён ниже:

Максимальный балл за ДЗ после просрочки дедлайна





# Примеры визуализаций по исследованию и сегментации

# Исследование распределений данных

## Числовые показатели

Открыть источник данных

Доступно Источники данных Импорт

Фильтр

\_VA\_BANK\_DATA\_RUS\_ABT\_174...  
14.03.2020, 16:00:43 • sas

\_VA\_BANK\_DATA\_RUS\_ABT\_58B...  
14.03.2020, 14:18:45 • sas

AUDIT  
15.03.2020, 13:08:49 • sas.ops-agentsrv

BANK\_DATA\_FOR\_STUDENTS  
25.02.2020, 16:08:09 • sasdemo

BANK\_DATA\_RUS\_ABT  
01.03.2020, 20:33:39 • msdemo12

CAS  
15.03.2020, 13:35:40 • sas.ops-agentsrv

CAS\_NODE  
15.03.2020, 13:35:40 • sas.ops-agentsrv

CAS\_SYSTEM  
15.03.2020, 13:35:40 • sas.ops-agentsrv

\_VA\_BANK\_DATA\_RUS\_ABT\_58B859AB-5E4D-4A4F-AF35-F8F40BC322DD\_3\_GE571

Сведения

Образец данных

Профиль

Отчет 15.03.2020, 13:37:42

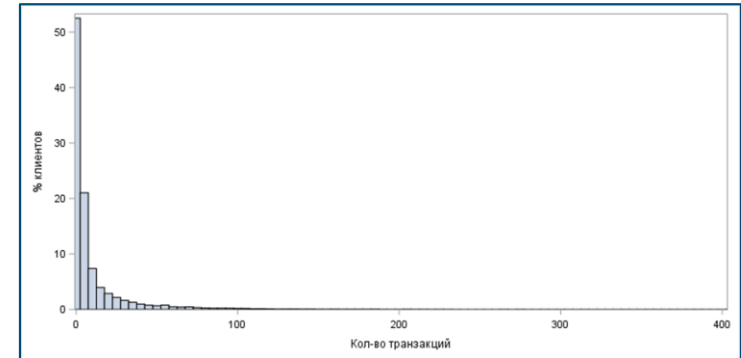
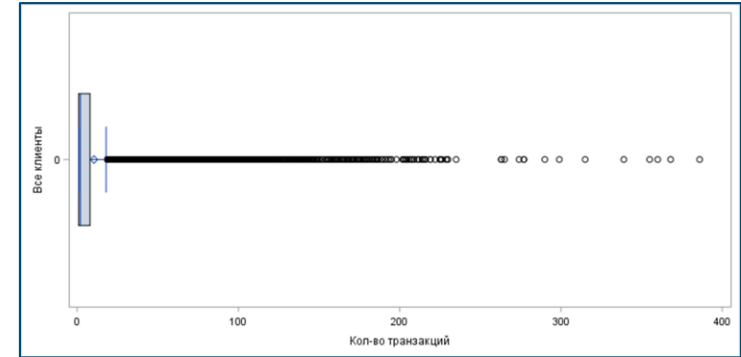
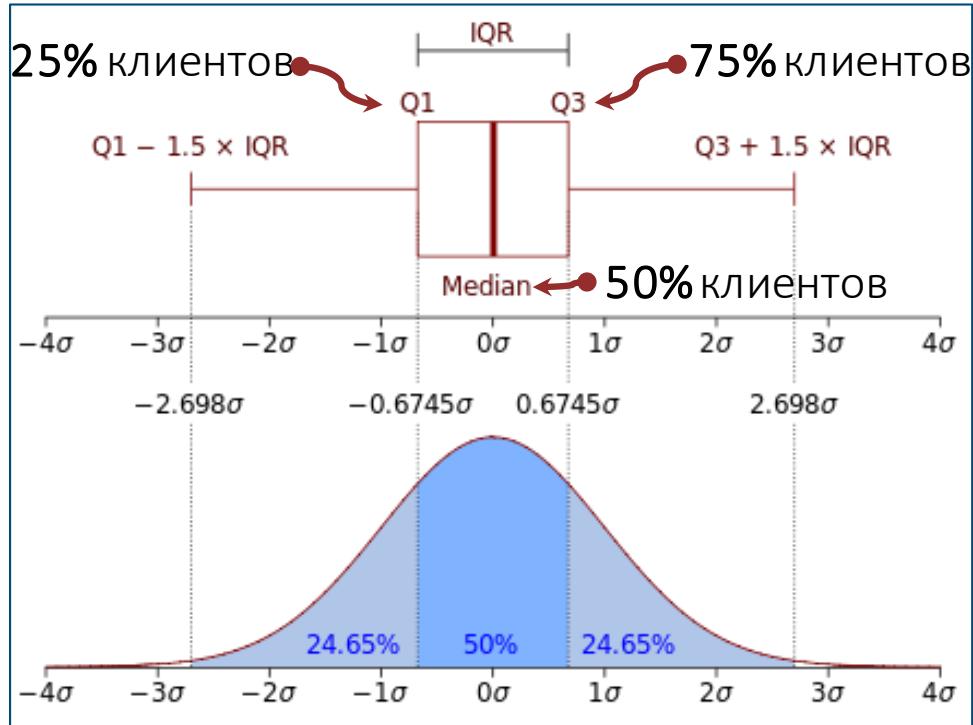
Отчет является текущим

Выполнить профилирование

Столбец	Уникальные	NULL	Пустой	Число обра...	Среднее	Медиана	Режим	Станда..
ID	100,00 % (10...)			1				
Segment	<0.01% (4)			4			Silver	
_PARTITION572	<0.01% (2)				0,70	1,00	1,00	0,46
Возраст	0,01 % (62)	6,80 % (713...)			53,79	54,00	51,00	13,20
Возрастная груп	<0.01% (4)			3			middle	
Время, проведе	<0.01% (39)	1,28 % (133...)			6,56	5,00	5,00	4,66
Доход	<0.01% (51)				50,35	50,00	50,00	5,44
Индикатор вла	<0.01% (2)			2			No	
Индикатор вла	<0.01% (2)			2			No	
Индикатор вла	<0.01% (2)			2			Yes	

# Исследование распределений данных

## Числовые показатели



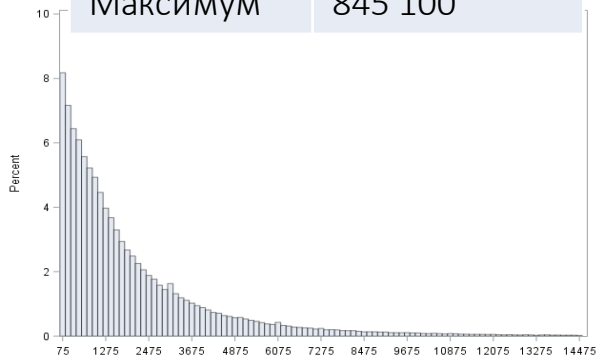


# Исследование распределений данных

## Распределения основных показателей

Сумма чека, руб.

Среднее	2 200
Минимум	0
P1%	4
P50%	1 450
P99%	24 856
Максимум	845 100



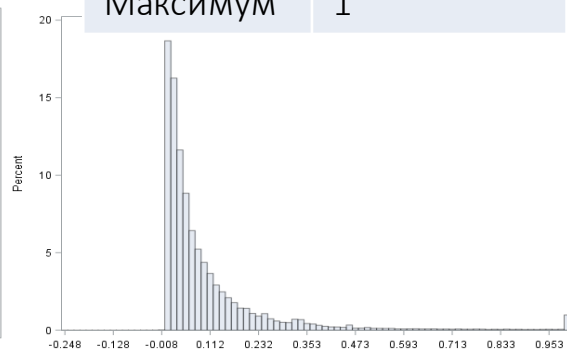
Скидка на чек, руб.

Среднее	156
Минимум	-1384
P1	0.01
P 50	14.15
P99	2 950
Максимум	250 000



Скидка на чек, доли

Среднее	0.1
Минимум	-0.35
P1	0.004
P 50	0.06
P99	0.97
Максимум	1



# Исследование распределений данных

## Странности в данных

Кол-во чеков в день у клиента  
(в дни, когда клиент посетил магазин)

Среднее	1.22
Минимум	1
P1%	1
P50%	1
P99%	3
Максимум	99

321 клиента, у которых хотя бы раз было более 3-х чеков в день:

- Доля от общей базы клиентов – 10%
- Доля по обороту – 15%
- Доля начисленной скидки – 16%

Количество товара qty в  
одном чеке

Среднее	8
Минимум	0
P1%	0.8
P50%	12
P99%	99
Максимум	2 359

1313 клиентов, у которых хотя бы раз было более 99 единиц товара в чеке:

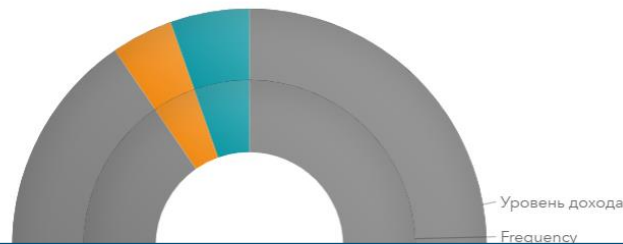
- Доля от общей базы клиентов – 6%
- Доля по обороту – 21%
- Доля начисленной скидки – 32%

# Пример графиков для исследования данных

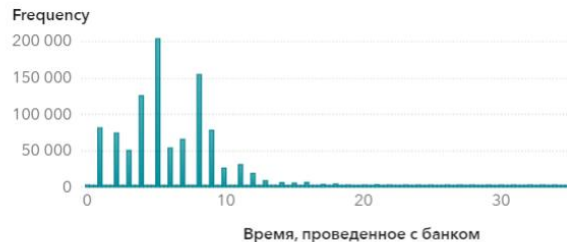
Транзакции за 12 мес по Возрастная группа



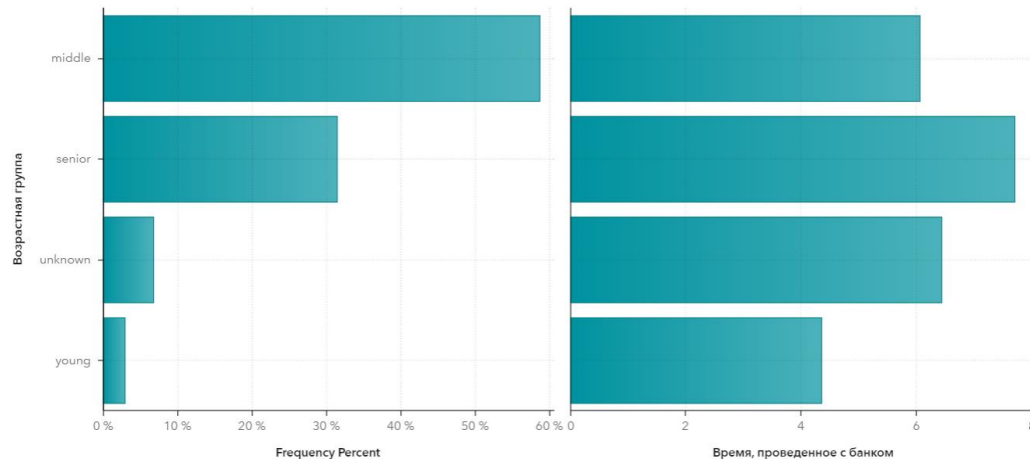
Уровень дохода, Frequency по Район



Frequency из Время, проведенное с банком



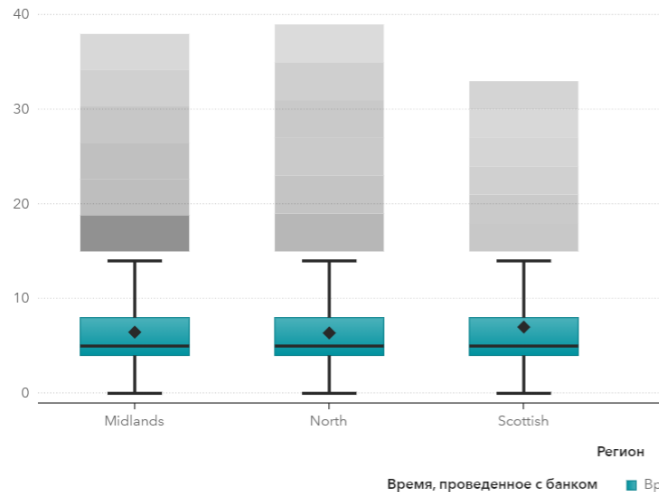
Frequency Percent, Время, проведенное с банком по Возрастная группа



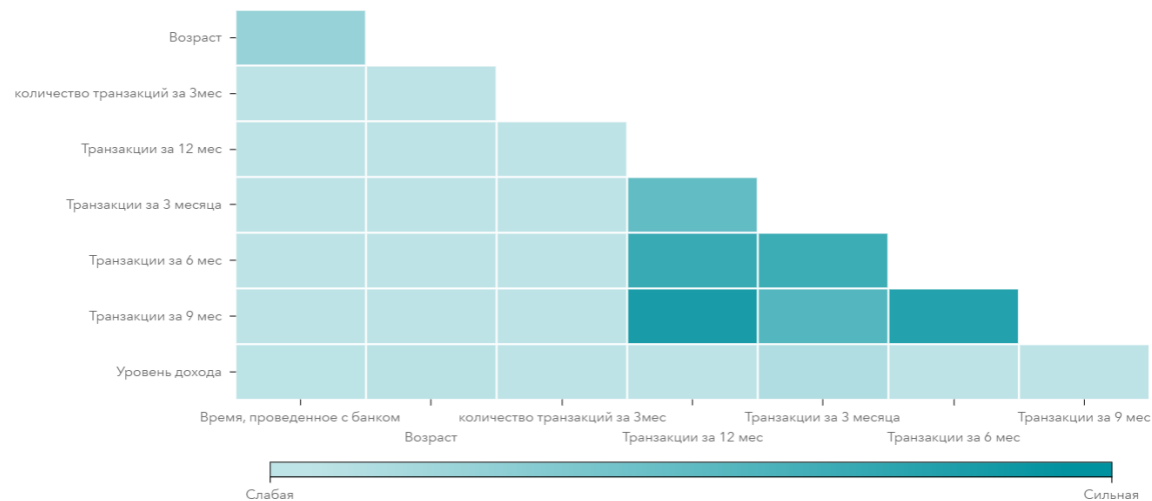
# Пример графиков для исследования данных

Время, проведенное с банком по Регион

Время, проведенное с банком

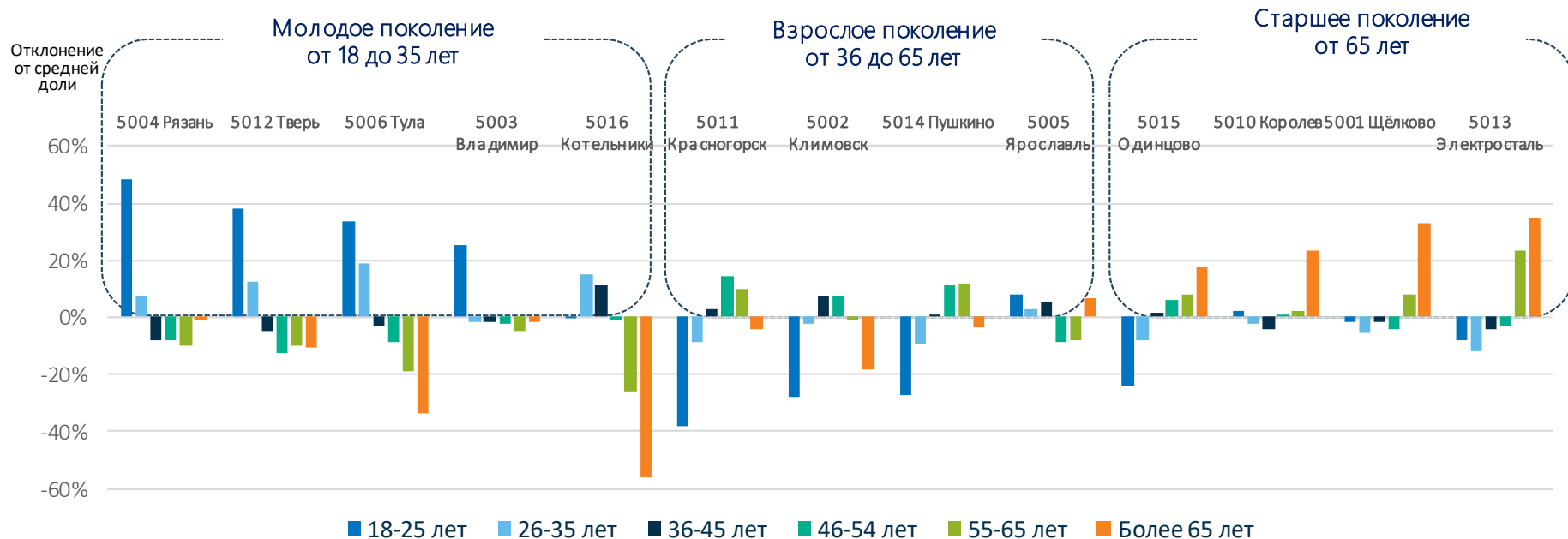


Корреляция выбранных показателей



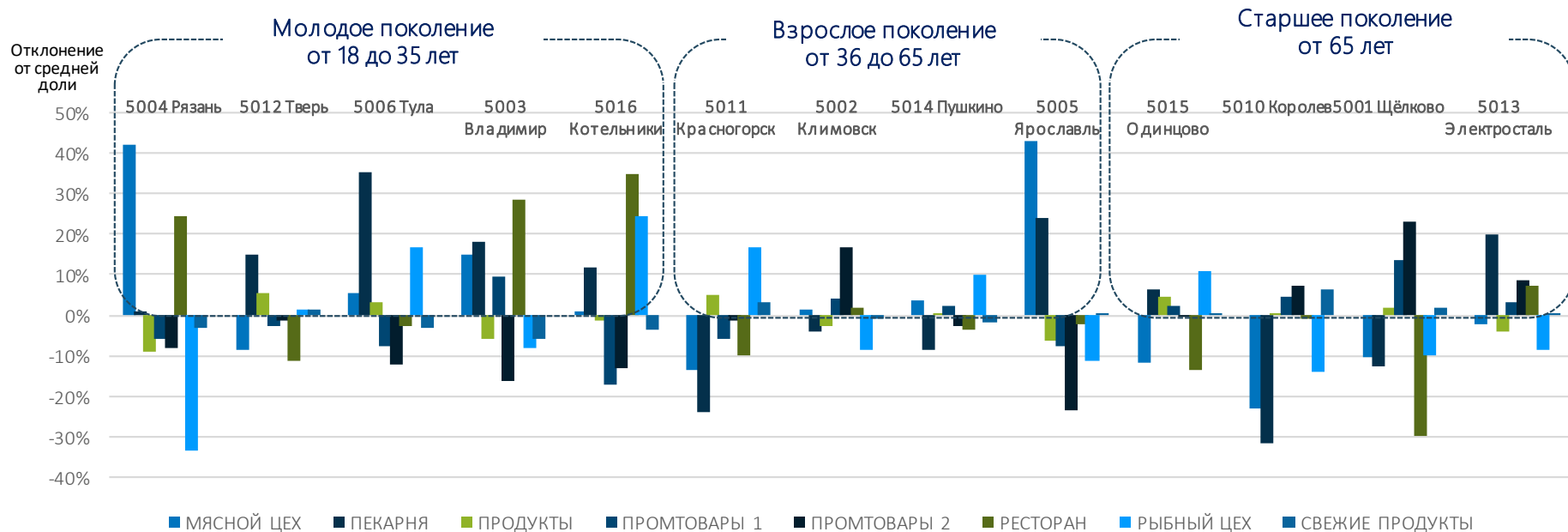
# Анализ различий покупателей в ГМ

## ГМ отличаются по возрастному профилю



# Анализ различий покупателей в ГМ

## ГМ отличаются по структуре спроса



# Сегментация: RFM-кластеризация

## Визуализация

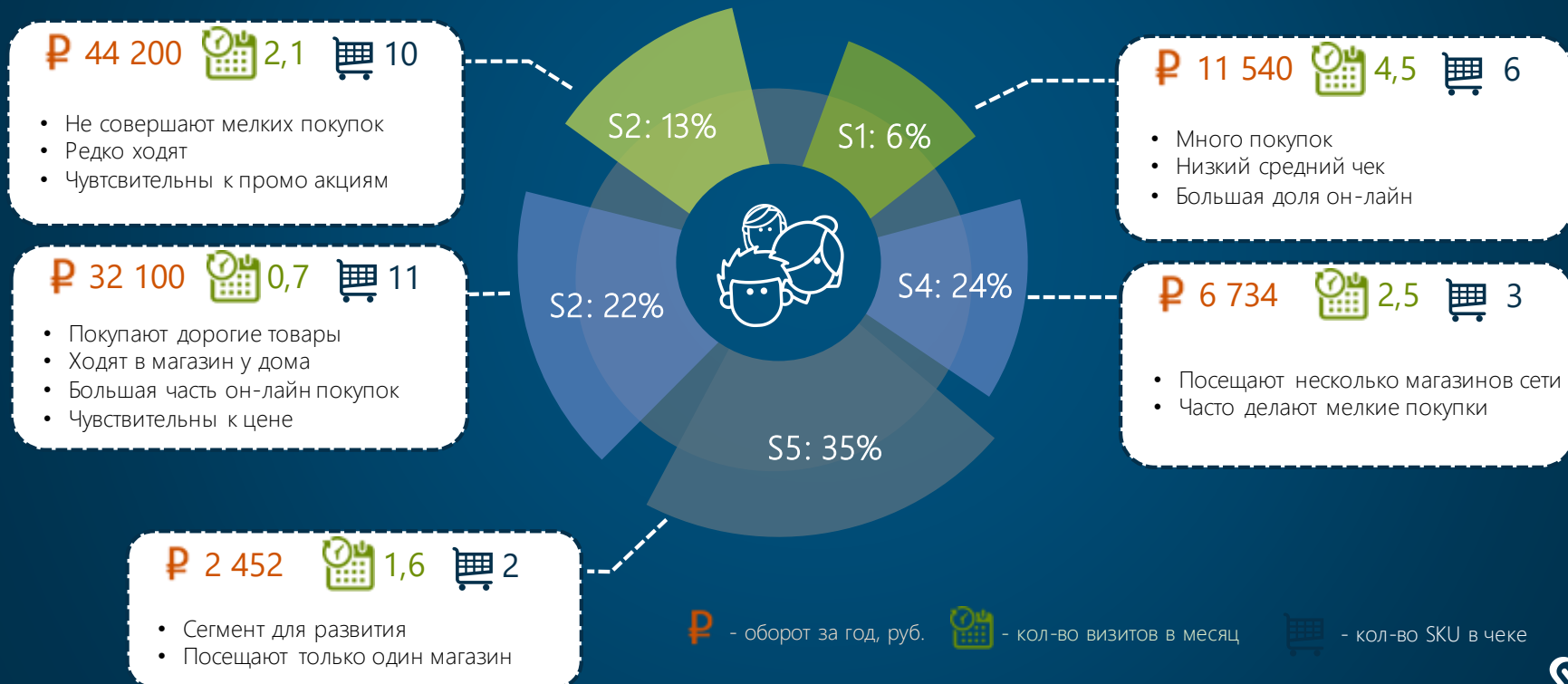
### RFM | СЕГМЕНТ «ДАВНО, НЕЧАСТО, МНОГО»



### RFM | СЕГМЕНТ «НЕДАВНО, ЧАСТО, МНОГО»

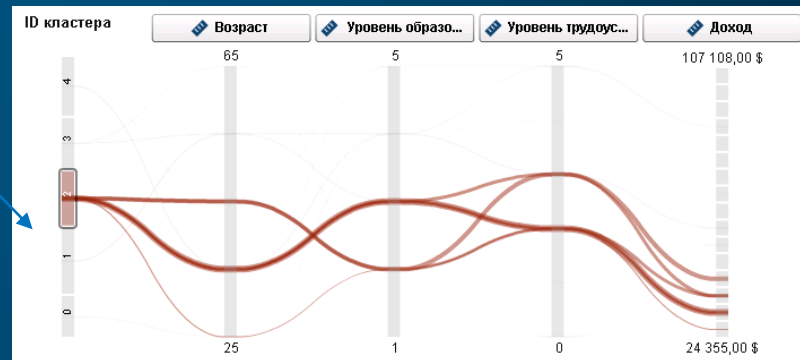
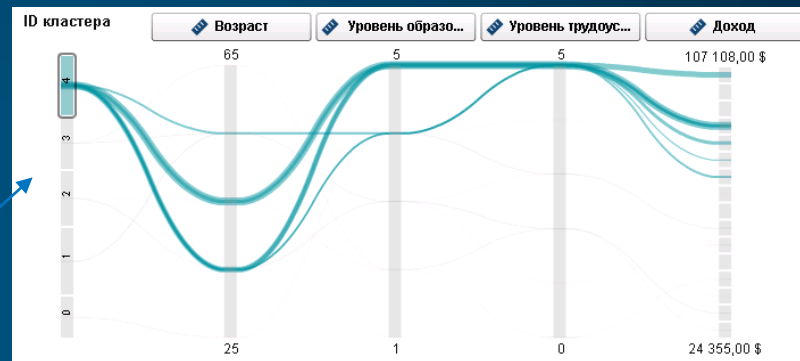
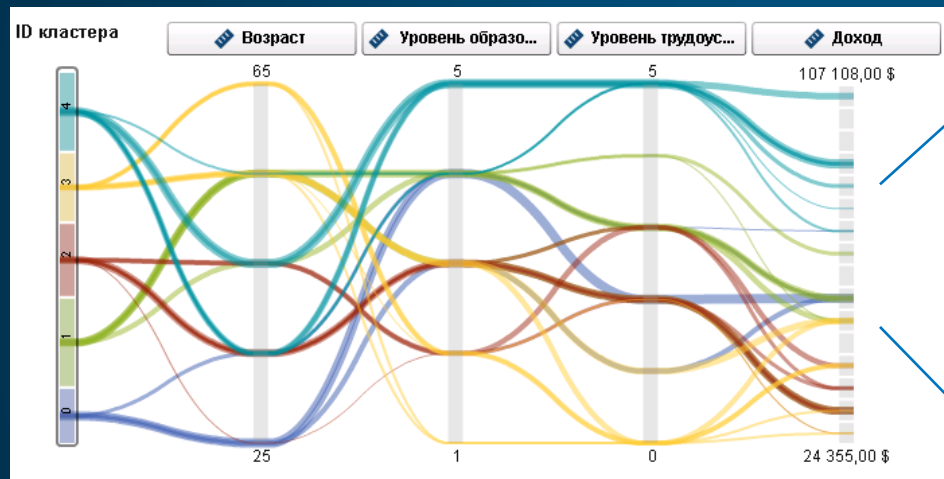


# Сегментация и мониторинг эффективности сегментов





# Сегментация без учителя в SAS VS/VA



Можно детально изучить каждый сегмент

# Пример профилирования одного кластера.

## Кластер 1 (ака средний рабочий)

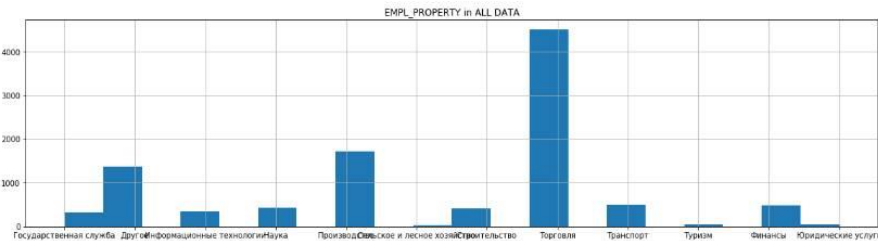
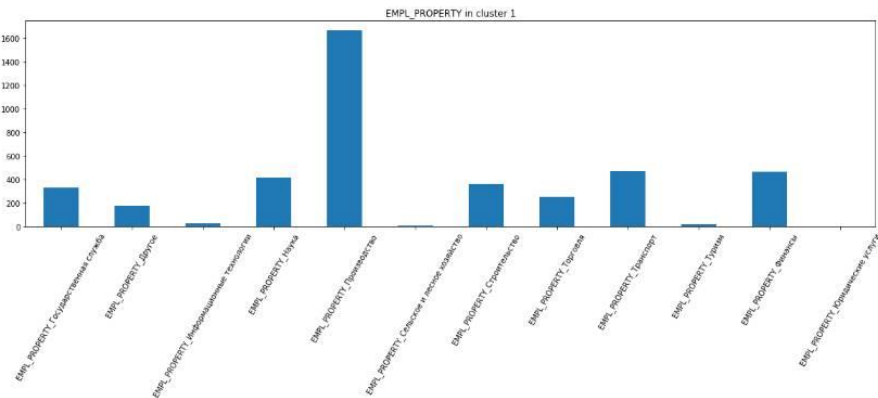
Это самый большой кластер. Из интересного тут это сильное большинство всех пользователей работают на производстве типа ООО. Также есть исключительно (!) только женат/замужем. =)

Можем смотреть на них как на стабильное большинство пользователей, базовый класс, стабильные люди в браке.

```
In [721]: print('Размер кластера:', len(kmeans.labels_[kmeans.labels_ == 1]))
```

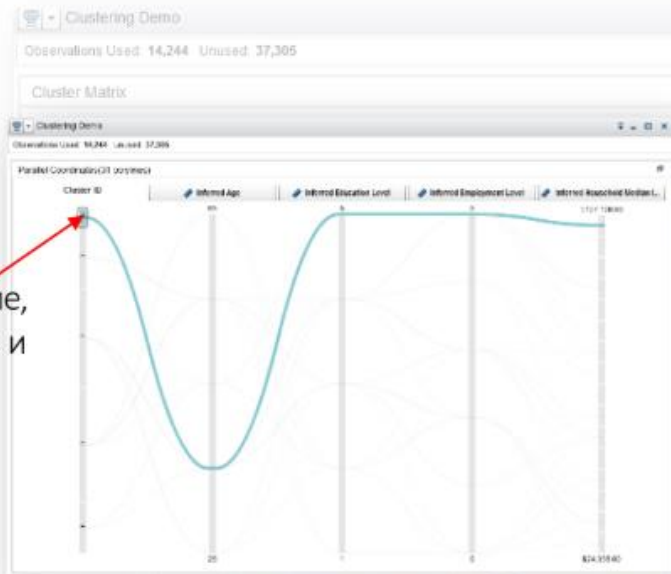
Размер кластера: 4201

```
In [703]: get_hists(kmeans, 1,
                  ['EMPL_PROPERTY', 'EMPL_FORM', 'FAMILY_STATUS'])
```

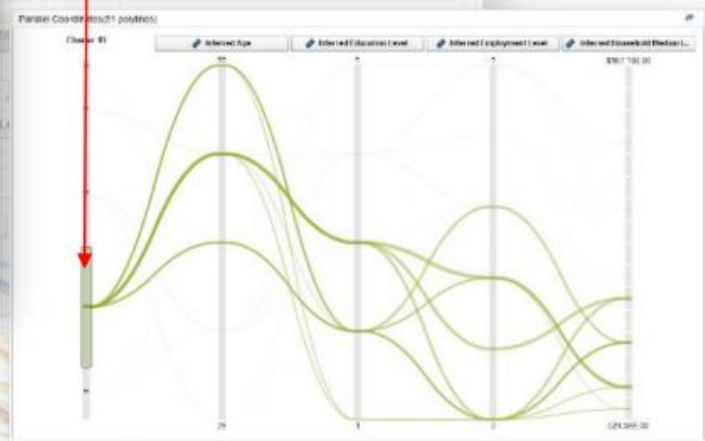


# Пример визуализации сегментации без учителя

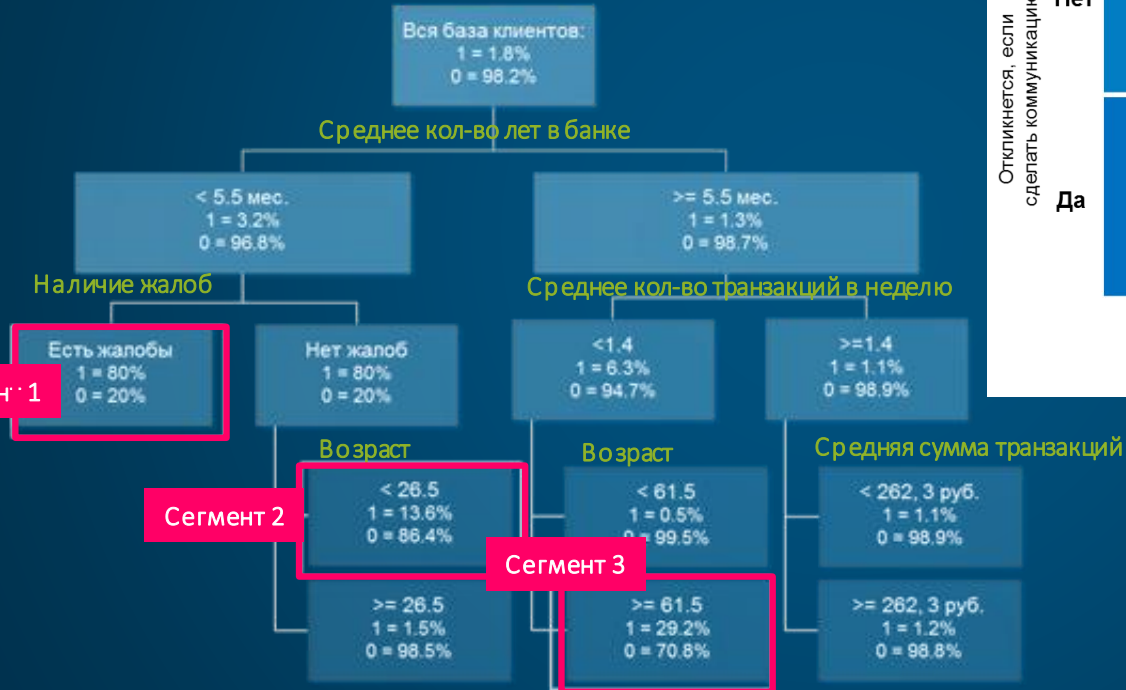
Посетители в этом кластере среднего возраста, имеют высшее образование, высокие должности и доход



Посетители в этом кластере старшего возраста, имеют различное образование, на пенсии и не зарабатывают много денег



# Сегментация с учителем (дерево решений)



# Пример результата

## Сопроводительное письмо

### Часть 1

По каждой переменной набора данных были рассчитаны: число уникальных значений (Unique и Percent\_Unique), число нулевых значений (Zeros, Percent\_Zeros), число пропущенных значений (NaNs, Percent\_NaN). Результаты приведены в таблице ниже:

	VARIABLE	UNIQUE	PERCENT_UNIQUE	ZEROS	PERCENT_ZEROS	NANS	PERCENT_NANS
0	ID	10242	79,93	0	0	0	0
1	INCOME_BASE_TYPE	5	0,04	0	0	68	0,04
2	CREDIT_PURPOSE	10	0,08	0	0	0	0
3	INSURANCE_FLAG	3	0,02	3992	0	2	0
4	DTI	61	0,48	3	0	128	0,07
5	SEX	2	0,02	0	0	0	0
6	FULL_AGE_CHILD_NUMBER	9	0,07	6016	0	0	0
7	DEPENDANT_NUMBER	3	0,02	10205	0	0	0
8	EDUCATION	9	0,07	0	0	0	0
9	EMPL_TYPE	10	0,08	0	0	8	0
10	EMPL_SIZE	9	0,07	0	0	125	0,07
11	BANKACCOUNT_FLAG	6	0,05	6219	0	2369	1,23
12	Period_at_work	364	2,84	0	0	2369	1,23
13	age	41	0,32	0	0	2369	1,23
14	EMPL_PROPERTY	13	0,1	0	0	2369	1,23
15	EMPL_FORM	7	0,05	0	0	6289	3,27
16	FAMILY_STATUS	7	0,05	0	0	6289	3,27
17	max90days	24	0,19	1079	0	6344	3,3
18	max60days	18	0,14	1524	0	6344	3,3
19	max30days	17	0,13	1973	0	6344	3,3
20	max21days	14	0,11	2350	0	6344	3,3

Красным цветом в таблице выделены переменные, имеющие наибольшее число пропущенных значений. Так как общее число пропусков составляет более половины наблюдений, нецелесообразно выбросить из набора данных наблюдения, имеющие пропуски, так как это приведёт к значительному сокращению выборки. Поэтому было решено заменить пропуски на глобальную константу: -57. Такая замена выгодна с точки зрения последующего анализа по двум причинам: во-первых, отрицательная константа не будет смешиваться с прочими – положительными – числами при анализе и построении моделей; во-вторых, большая по модулю константа будет легко отличима на гистограммах, что упрощает визуальный анализ.

Общий анализ уникальных и нулевых значений не выявил каких-либо отклонений, требуется дополнительное исследование.

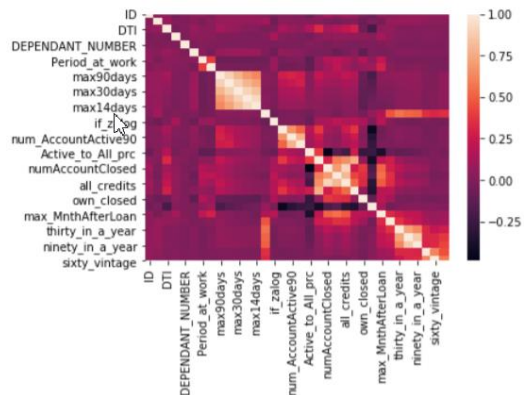
В таблице ниже приведены описательные статистики по численным переменным набора данных:

	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
ID	10242	1102424	59135,1	1000014	1051219	1102424	1153629	1204834
INSURANCE_FLAG	10240	0,61	0,49	0	0	1	1	1
DTI	10114	0,39	0,14	0	0,28	0,4	0,49	0,59
FULL_AGE_CHILD_NUMBER	10242	0,57	0,81	0	0	0	1	21
DEPENDANT_NUMBER	10242	0	0,07	0	0	0	0	2
BANKACCOUNT_FLAG	7873	0,39	0,88	0	0	0	0	4
PERIOD_AT_WORK	7873	64,52	65,09	4	19	43	85	460
AGE	7873	36,06	8,6	23	29	34	42	63

# Пример результата

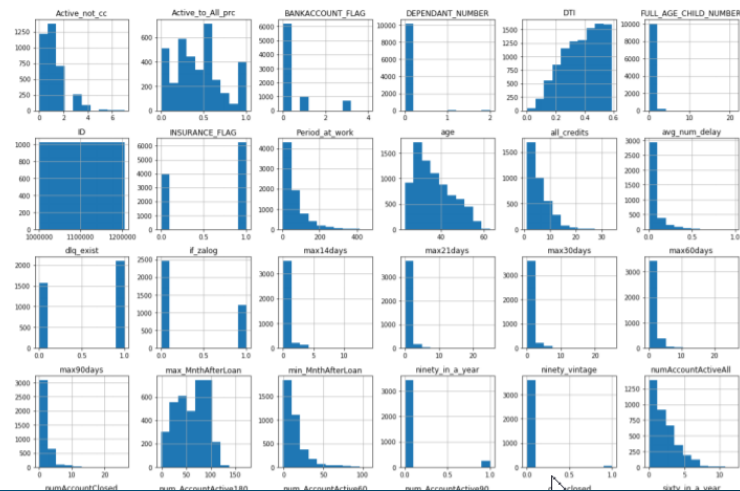
При анализе описательных статистик в данных была выявлена следующая странность: минимальное значение переменных MIN\_MNTHAFTERLOAN и MAX\_MNTHAFTERLOAN составляет -1, хотя эти переменные показывают время, прошедшее с момента выдачи последнего и первого кредитов соответственно. Так как неясно, несёт ли данное число содержательную информацию, или является показателем, например, пропущенного значения, указанные переменные было решено исключить из выборки.

По численным переменным была построена корреляционная карта:



Карта показывает, что в основном, между переменными наблюдается низкая корреляция, а потому линейно-регрессионный анализ в данном случае может быть неуместен.

На рисунке ниже приведены гистограммы по численным переменным (с опущенными пропусками) набора данных:

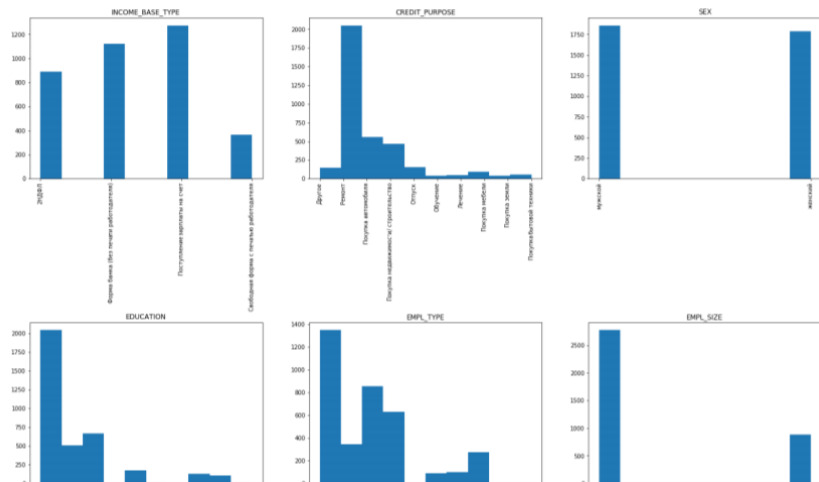




# Пример результата

Из визуального анализа гистограмм можно заметить, что в выборке присутствует, в основном, молодое поколение (до 40 лет), и отсутствуют пожилые люди (после 60 лет) (AGE). Также в выборке достаточно много индивидов с высоким отношением долга к доходам (DTI). Наконец, велико число человек с низким числом активных кредитов и банковских счетов (ALL\_CREDITS и NUMACCOUNTSACTIVEALL), а также низким количеством дней работы (PERIOD\_AT\_WORK) и числом платежей за последний месяц (SUM\_OF\_PAYM\_MONTHS). Это наводит на мысль, что в данных должны проявиться как минимум следующие кластеры: молодые люди, ещё не бравшие кредитов; молодые люди, взявшие один кредит и имеющие низкие доходы, так, что соотношение долга к доходам значительно; люди среднего возраста, имеющие активную кредитную историю. Можно отметить, что соотношение бинарных переменных непропорционально, а потому они не являются подходящими для сегментации признаками.

На рисунке ниже приведены гистограммы категориальных переменных набора данных:



Анализ диаграммы подтверждает выводы из исследования гистограмм: переменные INCOME\_BASE\_TYPE и SEX не имеют выбросов и распределены пропорционально, что сохраняет принцип однородности-разнородности, так как нет «скошенности» в одну сторону одной категории, и категории чётко различимы в данных.

В финальную витрину не вошли переменные без описания или технического характера: ID, SIXTY\_IN\_A\_YEAR, NINETY\_IN\_A\_YEAR, THIRTY\_VINTAGE, SIXTY\_VINTAGE, NINETY\_VINTAGE – а также переменные со странностями: MIN\_MNTHAFTERLOAN, MAX\_MNTHAFTERLOAN.

## Выводы по Части 1:

- В данных достаточно много пропущенных значений, выбрасывать их нецелесообразно. Пропущенные значения заменяются константой -57.
- Переменные MIN\_MNTHAFTERLOAN и MAX\_MNTHAFTERLOAN имеют недопустимые значения.
- Переменные слабо коррелируют между собой, линейно-регрессионный анализ неуместен.
- Для кодировки пропущенных значений в данных также присутствует строковая метка \*n.a.\*

С учётом всех корректировок была составлена финальная витрина.

# Пример результата

## Часть 2

### Сегментация с учителем (Decision Tree Classifier).

Чтобы сделать сегментацию осмысленной, для построения решающего дерева было решено отобрать небольшой набор переменных, легко интерпретируемых впоследствии. На основании предварительного анализа данных, было решено взять следующие переменные в качестве признаков: DTI, SEX, AGE, PERIOD\_AT\_WORK, SUM\_OF\_PAYM\_MONTHS, NUMACCOUNTACTIVEALL, ALL\_CREDITS – и добавить следующие переменные в качестве контрольных (из экономической логики): DLQ\_EXIST, CREDIT\_PURPOSE, EDUCATION. Так как в качестве предсказываемой переменной нужно брать переменную с примерно пропорциональным распределением значений (чтобы у классификатора не было соблазна всегда предсказывать только доминирующий класс), было решено в качестве таргета взять INCOME\_BASE\_TYPE. Глубина дерева была ограничена 4, чтобы сделать результаты его обучения читабельными.

Обученное решающее дерево выделило следующие сегменты (дерево также научилось предсказывать пропуски в данных, см. Notebook):

- Индивиды с низким отношением долгов к доходам, имеющие высшее образование, моложе 32 лет, работающие более трёх месяцев.
- Индивиды с низким отношением долгов к доходам, имеющие высшее образование, старше 32 лет, работающие менее двух месяцев.
- Индивиды с низким отношением долгов к доходам, имеющие высшее образование, старше 32 лет, работающие более двух месяцев.
- Индивиды с низким отношением долгов к доходам, имеющие среднее образование.
- Индивиды с низким отношением долгов к доходам, имеющие незаконченное высшее образование или учёную степень, работающие менее трёх месяцев.
- Индивиды с низким отношением долгов к доходам, имеющие незаконченное высшее образование или учёную степень, работающие более трёх месяцев.
- Индивиды со средним отношением долгов к доходам (менее 0.4), заплатившие менее 50 тыс. за последний месяц.
- Индивиды со средним отношением долгов к доходам (менее 0.4), заплатившие более 50 тыс. за последний месяц, работающие менее трёх месяцев.
- Индивиды со средним отношением долгов к доходам (менее 0.4), заплатившие более 50 тыс. за последний месяц, работающие более трёх месяцев.
- Индивиды с высоким отношением долгов к доходам (более 0.4), заплатившие менее 110 тыс. за последний месяц, по которым имеются данные по числу взятых кредитов.

Зелёным цветом в списке выделены интерпретируемые сегменты. Ниже приведён комментарий относительно отвергнутых сегментов:

- Индивиды с низким отношением долгов к доходам, имеющие высшее образование, старше 32 лет, работающие менее двух месяцев. – *Сомнительно, что индивиды старше 32 лет и имеющие высшее образование работали менее двух месяцев.*
- Индивиды с низким отношением долгов к доходам, имеющие среднее образование. – *Слишком широкий сегмент, недостаточно информации для выделения.*
- Индивиды с низким отношением долгов к доходам, имеющие незаконченное высшее образование или учёную степень, работающие менее трёх месяцев и Индивиды с низким отношением долгов к доходам, имеющие незаконченное высшее образование или учёную степень, работающие более трёх месяцев. – *Сомнительно, что в один сегмент попали люди, имеющие незаконченное высшее образование и учёную степень.*
- Индивиды со средним отношением долгов к доходам (менее 0.4), заплатившие более 50 тыс. за последний месяц, работающие менее трёх месяцев. – *Сомнительно, что индивиды, работающие менее трёх месяцев, способны совершить крупный платёж.*
- Индивиды с высоким отношением долгов к доходам (более 0.4), заплатившие менее 110 тыс. за последний месяц, по которым имеются данные по числу взятых кредитов. – *Слишком широкий сегмент, недостаточно информации для выделения.*

На основе интерпретируемых сегментов можно выделить следующие профили клиентов:

Плоскость 1:

1. Молодые люди (до 30 лет), недавно окончившие ВУЗ и работающие в данный момент, не имеющие активной кредитной истории.
2. Взрослые люди (старше 30 лет), имеющие высшее образование и работающие, но не имеющие активной кредитной истории.

Плоскость 2:

3. Индивиды, имеющие небольшой кредит, платежи по которому в месяц составляют менее 50 тыс. (например, бытовая и персональная техника).
4. Работающие индивиды с высоким заработком, оплачивающие по кредиту более 50 тыс. в месяц (например, работники технологических профессий, оплачивающие ипотеку в Москве).
5. Индивиды, активно пользующиеся банковскими счетами, имеющие высокие доходы и совершающие платежи более 110 тыс. в месяц (например, работники крупных банков, оплачивающие кредит на покупку дома).



# Пример результата

## Рекомендации:

### Кампания в Плоскости 1:

- Цель: привлечь клиентов в банк, запустить активное взаимодействие с банковскими сервисами.
- Особенности: различие клиентов только по возрасту; обе категории клиентов имеют высшее образование и работают в данный момент.
- Потенциальное предложение: кредит на универсальные нужды (обеим возрастным категориям актуальны покупки бытовой техники, персональных устройств, поездок).

### Кампания в Плоскости 2:

- Цель: обеспечить условия для сохранения клиентов выделенных категорий.

- Особенности: клиенты разделяются по типам кредита: небольшая покупка – средняя покупка – крупная покупка.
- Потенциальное предложение: три вида кредита: небольшая – средняя – крупная сумма. Возможность включения дополнительных опций (например, к крупным кредитам на покупку недвижимости предлагать скидку на страховку данной недвижимости в первые несколько лет). Также требуется дополнительная проверка заёмщиков третьего типа, так как они имеют большое значение отношения долга к доходу, то есть могут являться потенциальными дефолтерами.

## Сегментация по бизнес-правилам.

Бизнес-задача: таргетировать различные возрастные группы в зависимости от их предпочтений.

На основе предварительного анализа данных и для обеспечения сопоставимости с результатами сегментации на основе деревьев, были введены следующие бизнес-правила (пороги выбраны на основании исследования описательных статистик и уникальных значений переменных):

1. По возрасту:
  - a. Молодое поколение: 18-30 лет.
  - b. Взрослое поколение: 31-63 года.
2. По цели взятия кредита:
  - a. Покупка: покупка автомобиля, недвижимости, мебели, земли, бытовой техники, строительство.
  - b. Персональные услуги: лечение, отпуск, обучение.
  - c. Ремонт.
  - d. Другое.

Таким образом, в соответствии с бизнес-правилами выделяется 8 предварительных сегментов. Дальнейший анализ необходим для поиска однородностей в сегментах и их возможного укрупнения, а также для получения дополнительной информации о сегментах. Было рассмотрено распределение на сегментах тех же переменных, что использовались при построении решающего дерева (см. Notebook).

### Переменная DTI:

- По всем сегментам, кроме Взрослые – Другое данная переменная либо равна 0, либо данные отсутствуют. В сегменте Взрослые – Другое переменная принимает значения от 0 до 0.6.

### Переменная SEX:

- Во всех сегментах, кроме Молодые – Персональные услуги и Взрослые – Персональные услуги, преобладают мужчины. В указанных сегментах преобладают женщины.

### Переменная NUMACCOUNTACTIVEALL:

- В сегментах Молодые – Персональные услуги, Молодые – Другое данные либо отсутствуют, либо переменная близка к 0. В прочих сегментах данные либо отсутствуют, либо переменная больше или равна 0.

### Переменная INCOME\_BASE\_TYPE:

- В сегментах Молодые – ... и Взрослые – Другое преобладает тип «Поступление зарплаты на счёт». В сегментах Взрослые – Персональные услуги и Взрослые – Ремонт – тип «Форма банка (без печати работодателя)». В сегменте Взрослые – Покупка – тип «2НДФЛ».

### Переменные PERIOD\_AT\_WORK, SUM\_OF\_PAYM\_MONTHS, ALL\_CREDITS, DLQ\_EXIST, EDUCATION:

- Примерно одинаковое распределение по всем сегментам.

Таким образом, можно выделить следующие профили клиентов:

1. Молодые – Покупка: получают заработную плату, не имеют крупных долгов.
2. Молодые – Персональные услуги: преобладают девушки, не имеющие активных счетов и крупных долгов, и получающие заработную плату, цель взятия кредита которых – обучение, отпуск, рекреация (например, образовательные кредиты или уход за собой).
3. Молодые – Ремонт: получают заработную плату, не имеют крупных долгов.
4. Молодые – Другое: получают заработную плату, не имеют активных счетов, не имеют крупных долгов.
5. Взрослые – Покупка: получают доход по форме «2НДФЛ», не имеют крупных долгов.
6. Взрослые – Персональные услуги: преобладают женщины, получающие доход по форме банка, не имеющие крупных долгов, цель взятия кредита которых – обучение, отпуск, рекреация (например, образовательные кредиты или уход за собой).
7. Взрослые – Ремонт: не имеют крупных долгов, получают доход по форме банка.
8. Взрослые – Другое: имеют крупные долги, получают заработную плату.

Удачно выделенные сегменты отмечены зелёным в списке выше. На основании этого профили либо перестроены, либо уточнены в следующих:

1. «Персональные услуги – Любого возраста». Особенность: преобладают девушки.
2. «Другое – Взрослые». Особенность: требуется тщательная проверка заёмщика, так как соотношение долга к зарплате велико. С одной стороны, это может говорить о высокой заработной плате и совершении крупных покупок (топ-менеджеры банков), с другой – о потенциальной возможности дефолта.
3. «Все прочие категории – Все прочие возрасты».

# Пример результата

## Рекомендации:

### Для профиля 1:

- Цель: стимулировать клиента на взятие новых кредитов.
- Потенциальное предложение: кредит с особыми условиями (более низкими ставками) на поездки, обучение, оздоровительные мероприятия. Дополнительные предложения (например, бесплатное создание валютной карты для страны поездки).

### Для профиля 2:

- Цель: обеспечить удержание платёжеспособных клиентов.
- Особенность: требуется особо тщательное отделение неплатёжеспособных клиентов.
- Потенциальное предложение: индивидуальные кредиты на крупные суммы. Вип-статус (обслуживание без очередей).

### Для профиля 3:

- Цель: привлечение новых клиентов.
- Потенциальное предложение: кредит на универсальные нужды: небольшие суммы, небольшие проценты. Необходимо обеспечить быстроту получения.

(продолжение на следующей странице)

## Преимущества и недостатки выбранных методов сегментации.

	РЕШАЮЩЕЕ ДЕРЕВО	БИЗНЕС-ПРАВИЛА
ПРЕИМУЩЕСТВА	<ul style="list-style-type: none"> <li>• Сегменты выделяются на основе обучения, личное мнение человека минимально.</li> <li>• На используемых данных сегменты хорошо интерпретируются.</li> <li>• Возможность рассмотрения результатов в нескольких плоскостях, что даёт больше возможностей для проведения кампаний.</li> </ul>	<ul style="list-style-type: none"> <li>• Простая идея (но технически утомительное исполнение!)</li> </ul>
НЕДОСТАТКИ	<ul style="list-style-type: none"> <li>• Результат сильно зависит от выбора признаков, зависимой переменной и случайности при обучении модели.</li> <li>• Дерево необходимо ограничивать в глубину, чтобы получить визуально читабельный результат.</li> </ul>	<ul style="list-style-type: none"> <li>• Требуется верное задание бизнес-правил.</li> <li>• Требуется значительное вмешательство человека, дополнительное изучение особенностей выделенных сегментов.</li> </ul>

На основании построенных сегментаций было решено выбрать вариант, предложенный решающим деревом, так как он более обоснован с точки зрения математики, более автоматизируем, так как предполагает меньшее участие человека, даёт более богатую схему сегментации.

## Выводы по Части 2:

- В качестве модели сегментации было решено взять модель, построенную на основе решающего дерева. Предлагается следующая кампания:

НАЗВАНИЕ ПРОДУКТА	ОПИСАНИЕ ПРОДУКТА	ЦЕЛЕВАЯ АУДИТОРИЯ	ОСОБЕННОСТИ
КРЕДИТ «УНИВЕРСАЛЬНЫЙ»	Покупка бытовой и личной техники, повседневный траты, поездки.	Работающие люди всех возрастов, имеющие высшее образование.	Гибкость и простота получения.
КРЕДИТ «ВХОДНОЙ»	Покупка бытовой и личной техники, автомобиля, строительство и ремонт.	Работающие люди всех возрастов, имеющие высшее образование.	Гибкость и простота получения.
КРЕДИТ «СРЕДНИЙ»	Покупка недвижимости, автомобиля, крупной техники.	Работающие люди всех возрастов, имеющие высшее образование.	Дополнительные услуги, например, страховка.
КРЕДИТ «КРУПНЫЙ»	Покупка недвижимости, автомобиля, крупной техники.	Руководители и топ-менеджеры различных компаний.	Дополнительные услуги, например, страховка.