**Домашняя работа по курсу**

**Методы и средства обработки больших данных**

Выполнила Клепинина К.Д.

Задание состоит из 2-ух частей:

1.Исследование данных и обработка данных для проведения последующей сегментации;

2.Составить профили клиентов на основе проведенных сегментаций

**Содержание**

[Часть 1 2](#_Toc100091956)

[Подготовка данных. 2](#_Toc100091957)

[Очистка данных. 7](#_Toc100091958)

[Предварительный анализ данных 9](#_Toc100091959)

[Часть 2. 13](#_Toc100091960)

[K-means 14](#_Toc100091961)

[Описание сегментов 20](#_Toc100091962)

[Решающее дерево 20](#_Toc100091963)

[Обоснование выбора методов 21](#_Toc100091964)

# **Часть 1**

## **Подготовка данных.**

Проверка на дубликаты показала, что дубликатов нет.

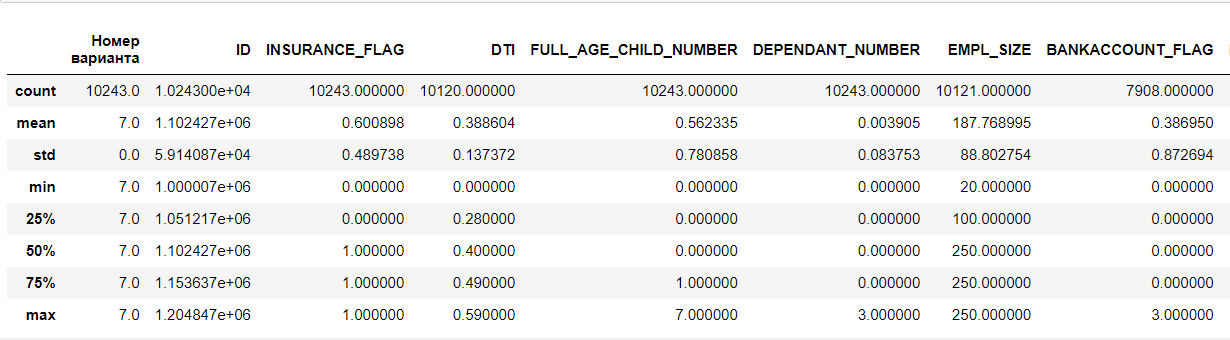
Среди значений таблицы есть неудобная для обработки переменная EMPL\_SIZE со значениями '>250', '>=50', '>=100', '>=200', '< 50', nan, '>100', '>=150'. Заменим эти значения на случайные в обозначенных интервалах (<50, от 50 до 100 и т.д.)

Рассмотрим количество уникальных и пропущенных значений и количество

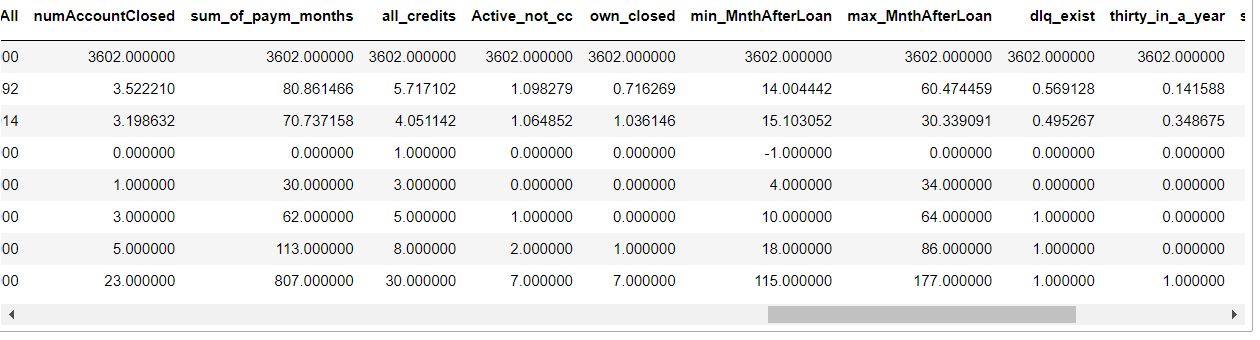
нулей для каждого значения.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | кол-во уникальных | кол-во нулей | доля нулей | кол-во пропущенных | доля пропущенных |
| Номер варианта | 1 | 0 | 0 | 0 | 0 |
| ID | 3963 | 0 | 0 | 0 | 0 |
| INCOME\_BASE\_TYPE | 4 | 0 | 0 | 0 | 0 |
| CREDIT\_PURPOSE | 10 | 0 | 0 | 0 | 0 |
| INSURANCE\_FLAG | 2 | 1541 | 38,88 | 0 | 0 |
| DTI | 57 | 0 | 0 | 0 | 0 |
| SEX | 2 | 0 | 0 | 0 | 0 |
| FULL\_AGE\_CHILD\_NUMBER | 6 | 2438 | 61,52 | 0 | 0 |
| DEPENDANT\_NUMBER | 4 | 3955 | 99,8 | 0 | 0 |
| EDUCATION | 9 | 0 | 0 | 0 | 0 |
| EMPL\_TYPE | 8 | 0 | 0 | 0 | 0 |
| EMPL\_SIZE | 759 | 0 | 0 | 1 | 0,03 |
| BANKACCOUNT\_FLAG | 3 | 3221 | 81,28 | 0 | 0 |
| Period\_at\_work | 264 | 0 | 0 | 1 | 0,03 |
| age | 38 | 0 | 0 | 0 | 0 |
| EMPL\_PROPERTY | 5 | 0 | 0 | 0 | 0 |
| EMPL\_FORM | 6 | 0 | 0 | 0 | 0 |
| FAMILY\_STATUS | 7 | 0 | 0 | 1 | 0,03 |
| max90days | 22 | 1041 | 26,27 | 60 | 1,51 |
| max60days | 20 | 1527 | 38,53 | 60 | 1,51 |
| max30days | 16 | 1936 | 48,85 | 60 | 1,51 |
| max21days | 15 | 2319 | 58,52 | 60 | 1,51 |
| max14days | 16 | 2499 | 63,06 | 60 | 1,51 |
| avg\_num\_delay | 1132 | 1538 | 38,81 | 375 | 9,46 |
| if\_zalog | 3 | 2400 | 60,56 | 361 | 9,11 |
| num\_AccountActive180 | 8 | 2561 | 64,62 | 361 | 9,11 |
| num\_AccountActive90 | 7 | 3073 | 77,54 | 361 | 9,11 |
| num\_AccountActive60 | 6 | 3263 | 82,34 | 361 | 9,11 |
| Active\_to\_All\_prc | 100 | 475 | 11,99 | 361 | 9,11 |
| numAccountActiveAll | 15 | 446 | 11,25 | 361 | 9,11 |
| numAccountClosed | 24 | 429 | 10,83 | 361 | 9,11 |
| sum\_of\_paym\_months | 325 | 14 | 0,35 | 361 | 9,11 |
| all\_credits | 30 | 0 | 0 | 361 | 9,11 |
| Active\_not\_cc | 9 | 1178 | 29,72 | 361 | 9,11 |
| own\_closed | 9 | 2058 | 51,93 | 361 | 9,11 |
| min\_MnthAfterLoan | 98 | 153 | 3,86 | 361 | 9,11 |
| max\_MnthAfterLoan | 129 | 11 | 0,28 | 361 | 9,11 |
| dlq\_exist | 3 | 1552 | 39,16 | 361 | 9,11 |
| thirty\_in\_a\_year | 3 | 3092 | 78,02 | 361 | 9,11 |
| sixty\_in\_a\_year | 3 | 3306 | 83,42 | 361 | 9,11 |
| ninety\_in\_a\_year | 3 | 3381 | 85,31 | 361 | 9,11 |
| thirty\_vintage | 3 | 3485 | 87,94 | 361 | 9,11 |
| sixty\_vintage | 3 | 3547 | 89,5 | 361 | 9,11 |
| ninety\_vintage | 3 | 3555 | 89,7 | 361 | 9,11 |

Рассмотрим среднее значение, медиана, стандартное отклонение, минимум, максимум по каждому столбцу данных:



Вызывает интерес минимальное значение переменной min\_MnthAfterLoan = -1, т.к. min\_MnthAfterLoan означает количество месяцев, прошедших со взятия последнего кредита.

Рассмотрим строки с этим значением переменной внимательнее, чтобы решить, как поступить с этим значением дальше. Есть несколько предположений о причине возникновения значения -1 в ячейке, в которой его не должно быть. Либо у этих клиентов никогда не было кредитов и значение ошибочно, в таком случае будем удалять эти строки, либо это планируемый кредит, т.е. кредит будет взят в следующем месяце, в этом случае заменим -1 на 0.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2997 | 3527 | 3550 | 6606 | 7709 | 9952 |
| Номер варианта | 7 | 7 | 7 | 7 | 7 | 7 |
| ID | 1059947 | 1070547 | 1071007 | 1132127 | 1154187 | 1199047 |
| INCOME\_BASE\_TYPE | Поступление зарплаты на счет | Поступление зарплаты на счет | Форма банка (без печати работодателя) | Форма банка (без печати работодателя) | 2НДФЛ | Форма банка (без печати работодателя) |
| CREDIT\_PURPOSE | Покупка земли | Другое | Покупка бытовой техники | Ремонт | Покупка недвижимости/ строительство | Ремонт |
| INSURANCE\_FLAG | 0 | 1 | 1 | 1 | 1 | 1 |
| DTI | 0,58 | 0,54 | 0,22 | 0,44 | 0,51 | 0,2 |
| SEX | женский | женский | мужской | женский | женский | женский |
| FULL\_AGE\_CHILD\_NUMBER | 0 | 0 | 1 | 0 | 0 | 2 |
| DEPENDANT\_NUMBER | 0 | 0 | 0 | 0 | 0 | 0 |
| EDUCATION | высшее | высшее | среднее | высшее | высшее | высшее |
| EMPL\_TYPE | специалист | вспомогательный персонал | специалист | специалист | рабочий | менеджер по продажам |
| EMPL\_SIZE | 250 | 250 | 250 | 250 | 250 | 250 |
| BANKACCOUNT\_FLAG | 1 | 0 | 0 | 0 | 0 | 0 |
| Period\_at\_work | 120 | 6 | 73 | 53 | 93 | 97 |
| age | 43 | 24 | 33 | 25 | 26 | 42 |
| EMPL\_PROPERTY | Другое | Другое | Другое | Торговля | Другое | Торговля |
| EMPL\_FORM | ЗАО | ООО | ООО | ООО | ОАО | ООО |
| FAMILY\_STATUS | женат / замужем | женат / замужем | женат / замужем | холост / не замужем | женат / замужем | разведен / разведена |
| max90days | 2 | 2 | 4 | 2 | 3 | 10 |
| max60days | 2 | 2 | 4 | 2 | 2 | 1 |
| max30days | 2 | 2 | 3 | 2 | 2 | 1 |
| max21days | 1 | 2 | 2 | 2 | 2 | 0 |
| max14days | 1 | 0 | 0 | 0 | 0 | 0 |
| avg\_num\_delay | 0,004587 | 0,040816 | 0 | 0,01 | 0 | 0,02834 |
| if\_zalog | 1 | 1 | 0 | 1 | 0 | 1 |
| num\_AccountActive180 | 1 | 2 | 4 | 1 | 2 | 2 |
| num\_AccountActive90 | 1 | 2 | 2 | 1 | 1 | 2 |
| num\_AccountActive60 | 1 | 2 | 2 | 1 | 1 | 2 |
| Active\_to\_All\_prc | 0,5 | 0,571429 | 0,8 | 0,5 | 0,571429 | 0,583333 |
| numAccountActiveAll | 5 | 4 | 4 | 5 | 4 | 7 |
| numAccountClosed | 5 | 3 | 1 | 5 | 3 | 5 |
| sum\_of\_paym\_months | 217 | 47 | 42 | 99 | 88 | 240 |
| all\_credits | 10 | 7 | 5 | 10 | 7 | 12 |
| Active\_not\_cc | 4 | 2 | 2 | 3 | 2 | 4 |
| own\_closed | 0 | 0 | 1 | 0 | 0 | 2 |
| min\_MnthAfterLoan | -1 | -1 | -1 | -1 | -1 | -1 |
| max\_MnthAfterLoan | 90 | 39 | 34 | 48 | 35 | 83 |
| dlq\_exist | 1 | 1 | 0 | 1 | 0 | 1 |
| thirty\_in\_a\_year | 0 | 0 | 0 | 0 | 0 | 1 |
| sixty\_in\_a\_year | 0 | 0 | 0 | 0 | 0 | 1 |
| ninety\_in\_a\_year | 0 | 0 | 0 | 0 | 0 | 1 |

Всего 6 клиентов с странным значением min\_MnthAfterLoan (таблица транспонирована для удобства восприятия). Можно заметить, что у каждого из клиентов есть как минимум 5 кредитов, т.е. теория об отсутствии кредитов у таких клиентов не подтвердилась. Для проверки второй теории посмотрим на значения других переменных, таких как количество активных счетов за последние 60 дней и количество запросов в бюро кредитных историй за последний месяц. У всех клиентов есть как минимум один активный за последние 60 дней счёт, но в последние 21-14 дней запросы в бюро кредитных историй были не у всех. Насколько я понимаю, запрос кредитной истории необходим для принятия решения о выдаче кредитов. Предположу возможность ситуации, в которой между запросом в бюро кредитных историй и моментом выдачи кредита клиенту проходит >21, но <30 дней. Оставим эти строки в исходном виде, посчитав, что числа несут содержательную информацию.

Также кажется интересным минимум равный 4 в столбце **Period\_at\_work,** где записано количество дней работы. Стаж работы для получения кредита в крупных банках должен быть не меньше 3-4 месяцев. Посмотрим на количество клиентов со стажем меньше месяца – таких найдено 2766, что составляет достаточно большой процент от всех значений базы данных, следовательно предположить ошибочное попадание данных сложно. Максимальное значение переменной **Period\_at\_work** равно 489, т.е. немногим больше года. Логичным кажется предположить ошибку в описании данных, скорее всего это значение стажа в месяцах, тогда максимальное значение переменной равно примерно 40 годам, что выглядит вполне реально при максимальном возрасте 62 года. Данные будем считать верными.

Проверим нечисловые переменные на наличие «странных» значений:



Нестандартных значений не обнаружено.

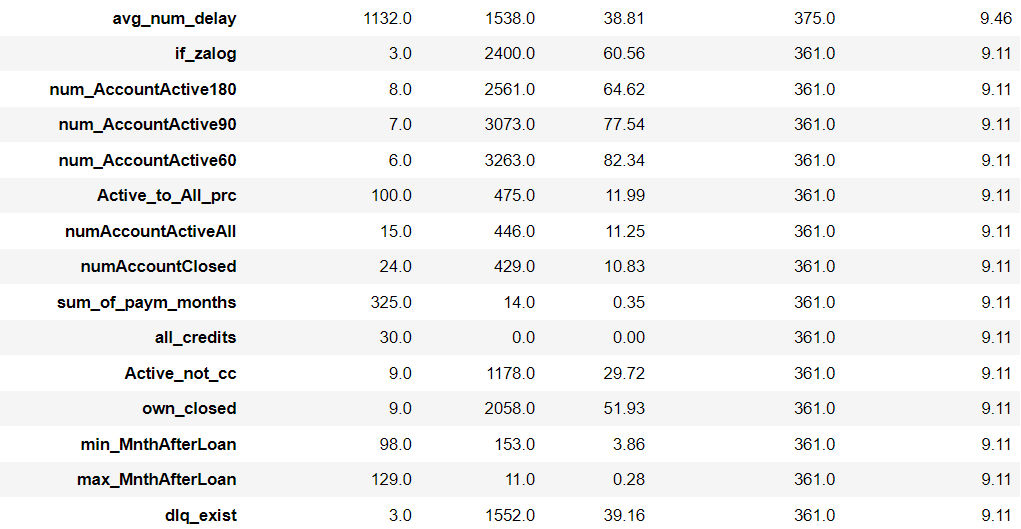
## **Очистка данных.**

Для каждой переменной было рассчитано число уникальных, нулевых и пропущенных значений:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | кол-во уникальных | кол-во нулей | доля нулей | кол-во пропущенных | доля пропущенных |
| Номер варианта | 1 | 0 | 0 | 0 | 0 |
| ID | 10243 | 0 | 0 | 0 | 0 |
| INCOME\_BASE\_TYPE | 5 | 0 | 0 | 63 | 0,62 |
| CREDIT\_PURPOSE | 10 | 0 | 0 | 0 | 0 |
| INSURANCE\_FLAG | 2 | 4088 | 39,91 | 0 | 0 |
| DTI | 59 | 1 | 0,01 | 123 | 1,2 |
| SEX | 2 | 0 | 0 | 0 | 0 |
| FULL\_AGE\_CHILD\_NUMBER | 8 | 6085 | 59,41 | 0 | 0 |
| DEPENDANT\_NUMBER | 4 | 10216 | 99,74 | 0 | 0 |
| EDUCATION | 8 | 0 | 0 | 0 | 0 |
| EMPL\_TYPE | 10 | 0 | 0 | 11 | 0,11 |
| EMPL\_SIZE | 474 | 0 | 0 | 122 | 1,19 |
| BANKACCOUNT\_FLAG | 4 | 6230 | 60,82 | 2335 | 22,8 |
| Period\_at\_work | 374 | 0 | 0 | 2337 | 22,82 |
| age | 39 | 0 | 0 | 2335 | 22,8 |
| EMPL\_PROPERTY | 13 | 0 | 0 | 2335 | 22,8 |
| EMPL\_FORM | 7 | 0 | 0 | 6280 | 61,31 |
| FAMILY\_STATUS | 7 | 0 | 0 | 6281 | 61,32 |
| max90days | 22 | 1041 | 10,16 | 6340 | 61,9 |
| max60days | 20 | 1527 | 14,91 | 6340 | 61,9 |
| max30days | 16 | 1936 | 18,9 | 6340 | 61,9 |
| max21days | 15 | 2319 | 22,64 | 6340 | 61,9 |
| max14days | 16 | 2499 | 24,4 | 6340 | 61,9 |
| avg\_num\_delay | 1132 | 1538 | 15,02 | 6655 | 64,97 |
| if\_zalog | 3 | 2400 | 23,43 | 6641 | 64,83 |
| num\_AccountActive180 | 8 | 2561 | 25 | 6641 | 64,83 |
| num\_AccountActive90 | 7 | 3073 | 30 | 6641 | 64,83 |
| num\_AccountActive60 | 6 | 3263 | 31,86 | 6641 | 64,83 |
| Active\_to\_All\_prc | 100 | 475 | 4,64 | 6641 | 64,83 |
| numAccountActiveAll | 15 | 446 | 4,35 | 6641 | 64,83 |
| numAccountClosed | 24 | 429 | 4,19 | 6641 | 64,83 |
| sum\_of\_paym\_months | 325 | 14 | 0,14 | 6641 | 64,83 |
| all\_credits | 30 | 0 | 0 | 6641 | 64,83 |
| Active\_not\_cc | 9 | 1178 | 11,5 | 6641 | 64,83 |
| own\_closed | 9 | 2058 | 20,09 | 6641 | 64,83 |
| min\_MnthAfterLoan | 98 | 153 | 1,49 | 6641 | 64,83 |
| max\_MnthAfterLoan | 129 | 11 | 0,11 | 6641 | 64,83 |
| dlq\_exist | 3 | 1552 | 15,15 | 6641 | 64,83 |
| thirty\_in\_a\_year | 3 | 3092 | 30,19 | 6641 | 64,83 |
| sixty\_in\_a\_year | 3 | 3306 | 32,28 | 6641 | 64,83 |
| ninety\_in\_a\_year | 3 | 3381 | 33,01 | 6641 | 64,83 |
| thirty\_vintage | 3 | 3485 | 34,02 | 6641 | 64,83 |
| sixty\_vintage | 3 | 3547 | 34,63 | 6641 | 64,83 |
| ninety\_vintage | 3 | 3555 | 34,71 | 6641 | 64,83 |

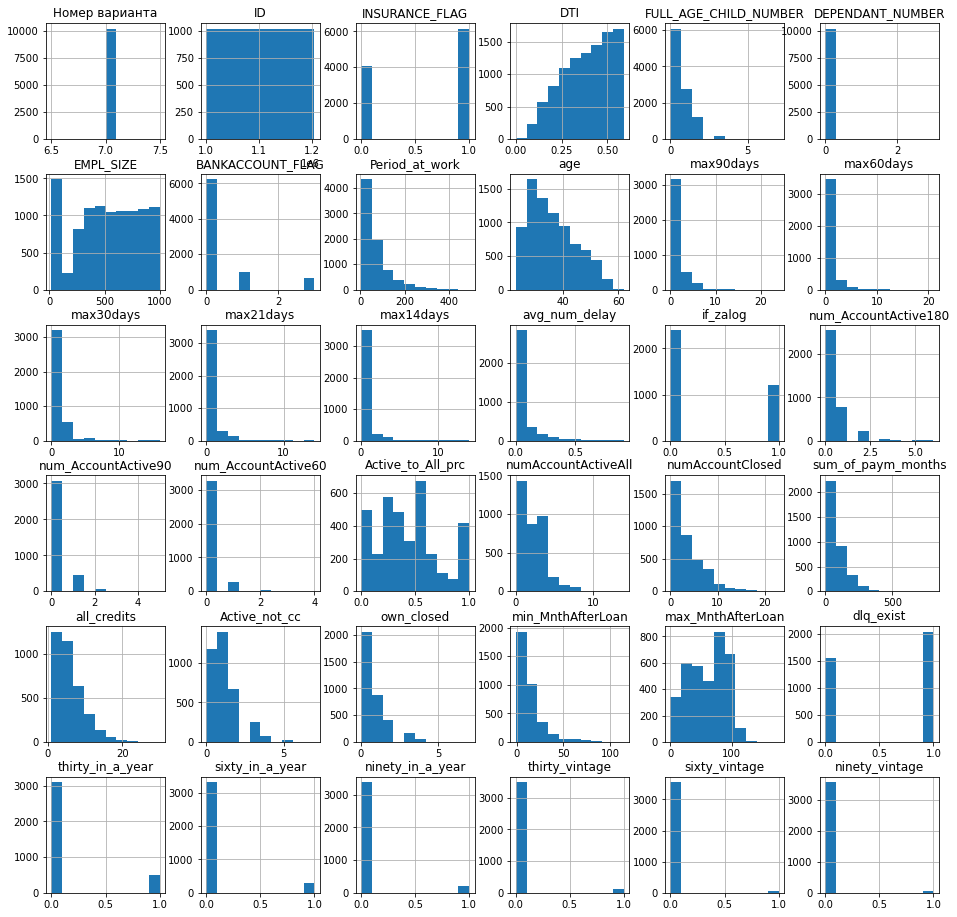
При очистке данных сохраним строки, в которых есть как минимум треть непустых значений. Пустые значения заполним медианными , категориальные переменные преобразуем в количественные при помощи OneHotEncoder.

После очистки данных максимальный процент пропущенных значений равен 9.46.



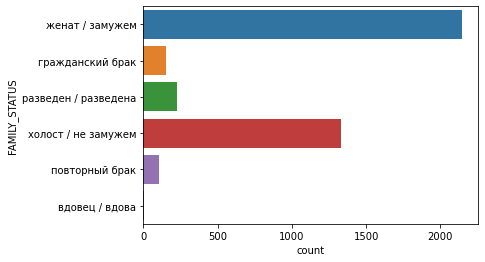
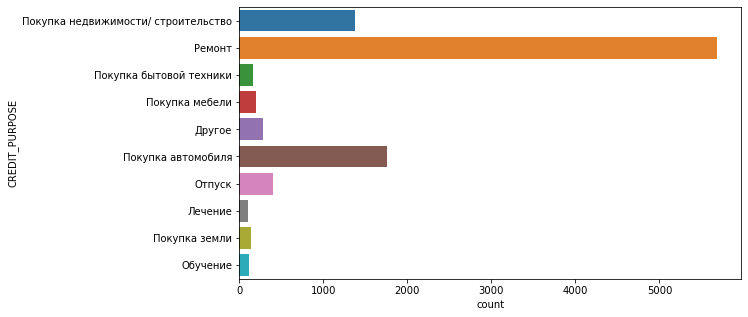
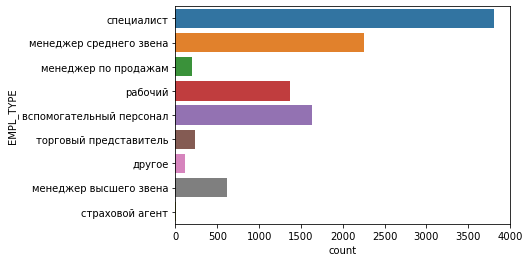
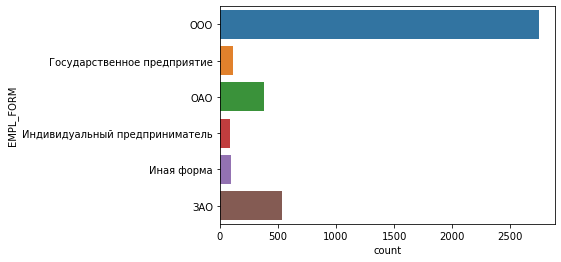
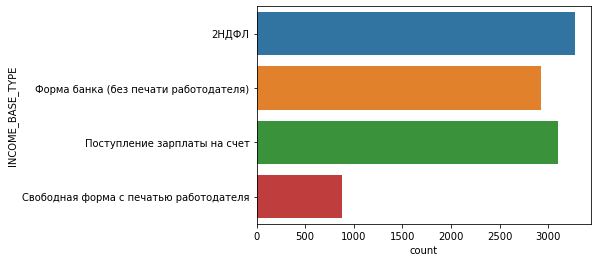
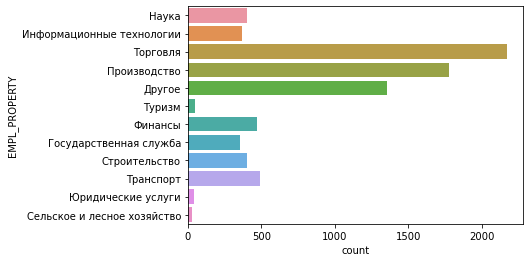
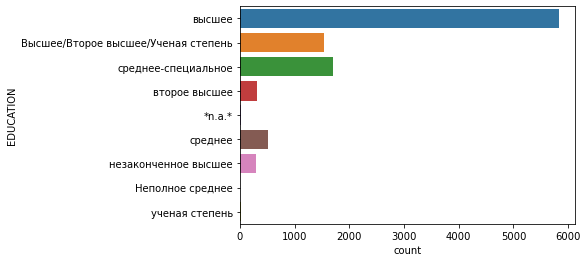
## **Предварительный анализ данных**

Построим гистограммы для всех некатегориальных переменных



Как видно из графиков, в выборке преобладают достаточно молодые клиенты, в основном от 30 до 40 лет и без детей или с маленькими детьми. Большая часть клиентов работает меньше 4 лет и получает зарплату от 250 тыс.руб. Большинство клиентов брали кредит за последний год, не имеют аккаунта в банке, просрочки по платежам, если происходят, меньше 30 дней.

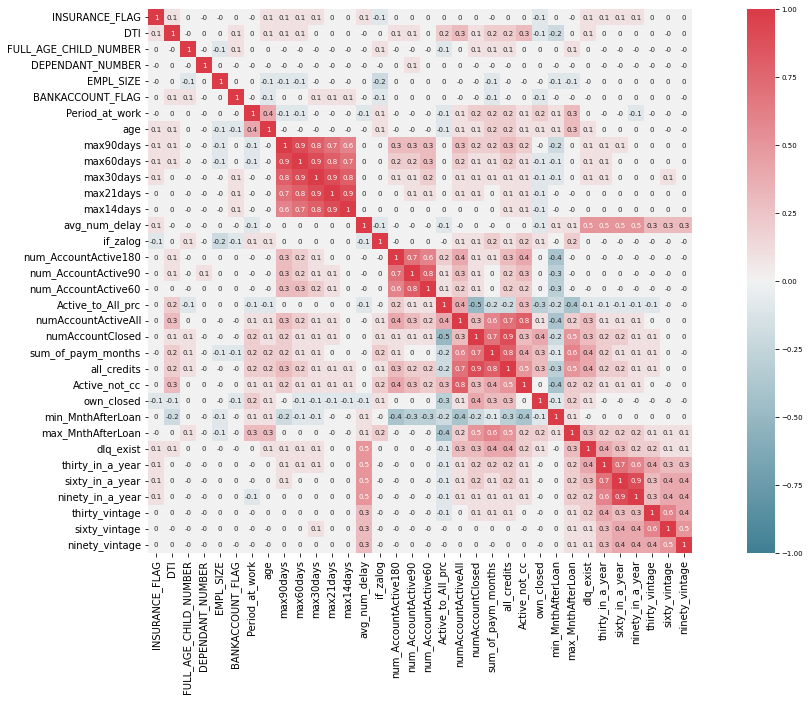
Также рассмотрим диаграммы по некатегориальным переменным.



**Часть 2.**

Для сегментации данных будем использовать метод k-means и дерево решений.

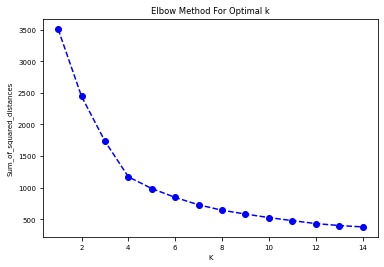
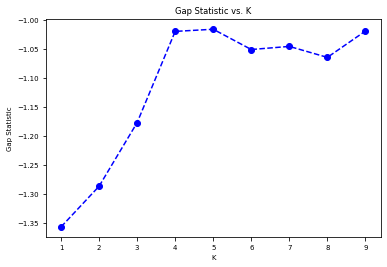
Сначала построим матрицу корреляций:



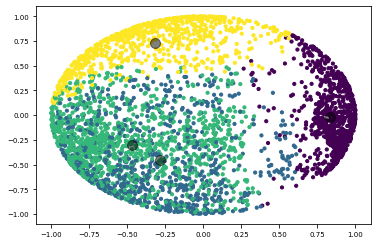
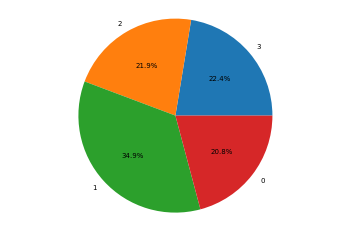
Как видно из матрицы, корреляция в целом достаточно слабая, но есть несколько очагов сильной корреляции между переменными, значения линейно зависимы, таких как, например, количество запросов в бюро кредитной истории за последние 14, 21, 30, 60 и 90 дней. Можно заменить существующие переменные, например, их отношением, но в рамках задания и используемых методов сегментации, в этом нет необходимости.

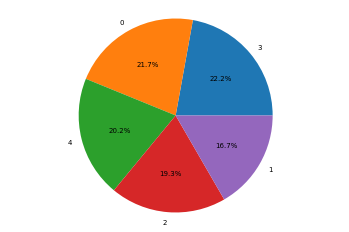
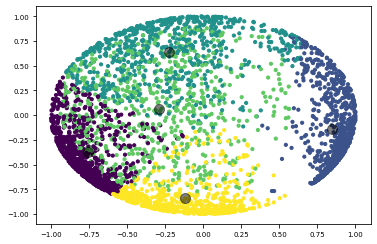
## **K-means**

Воспользуемся методом главных компонент для понижения размерности и определим оптимальное количество кластеров при помощи gap statistics и elbow method.

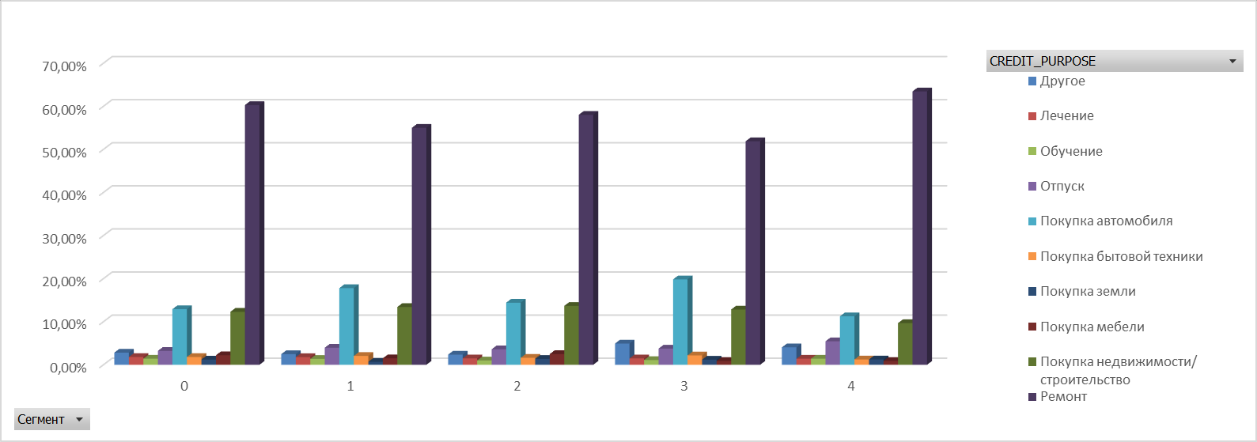
Исходя из графиков, предположим оптимальное количество кластеров равное 4. Однако, при выборе 5 кластеров, результат работы алгоритма выглядит лучше.

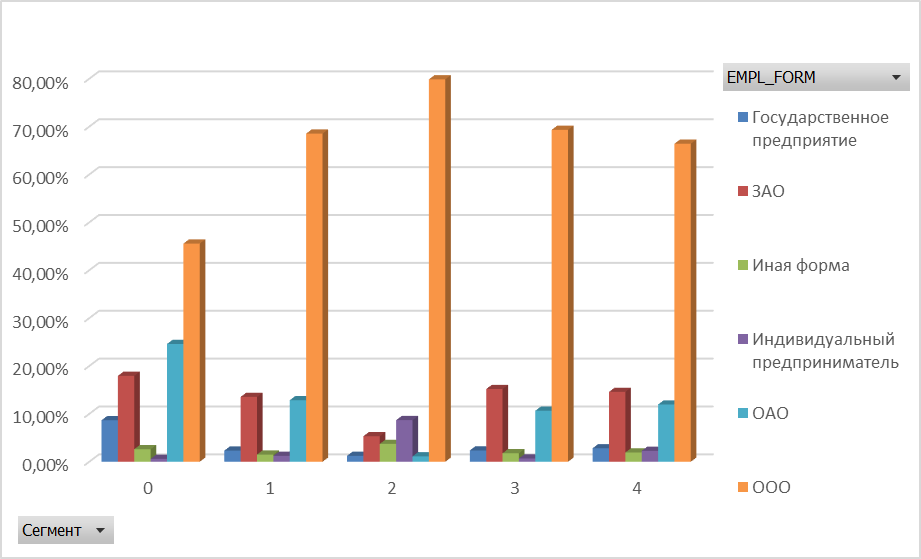
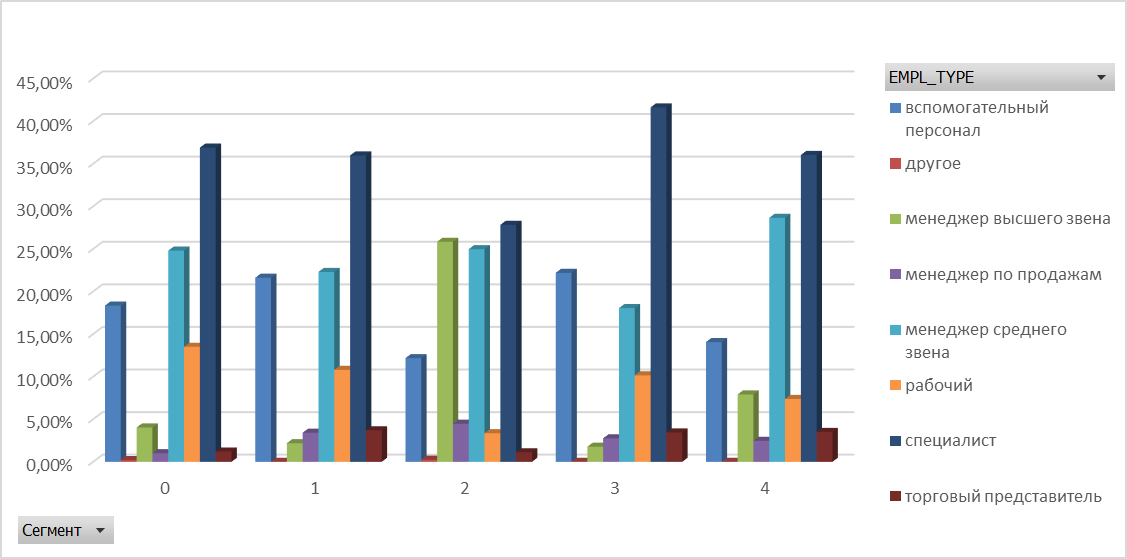


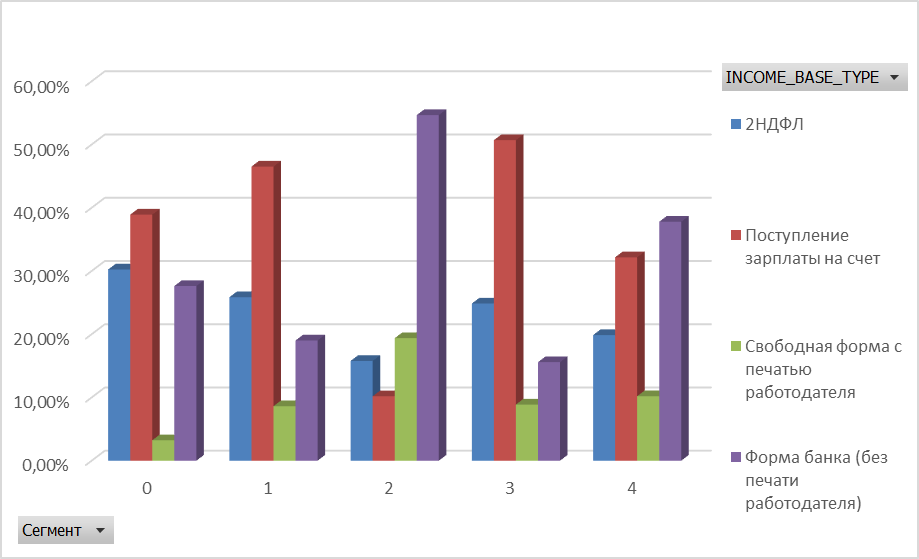
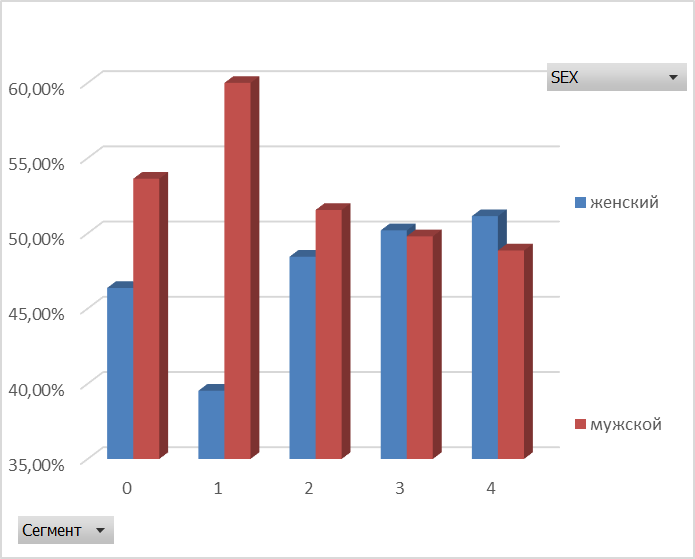
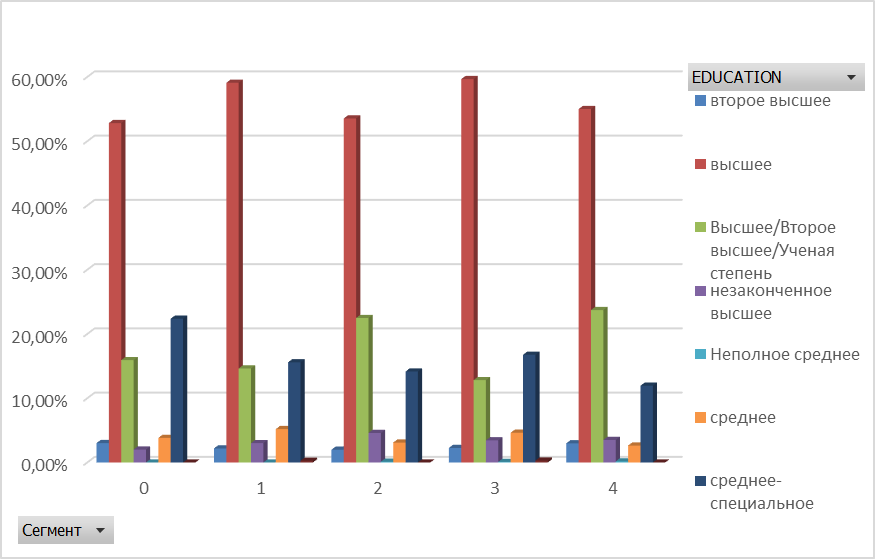
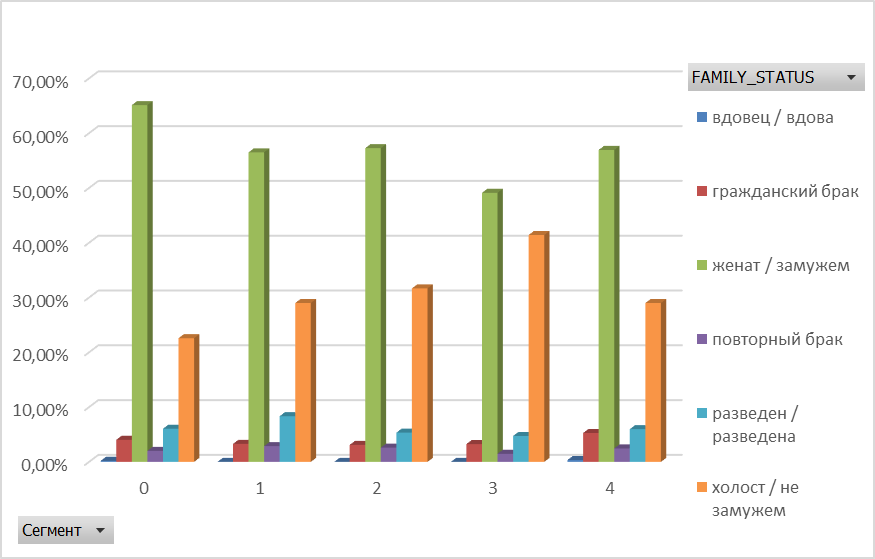
 

Результаты работы алгоритма

Теперь рассмотрим гистограммы признаков для сегментов:

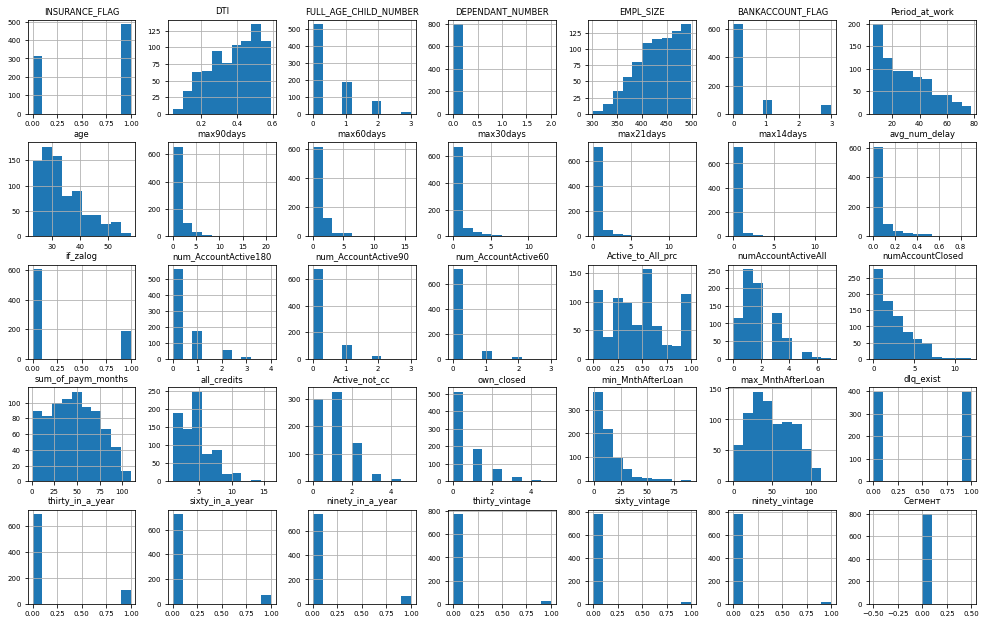




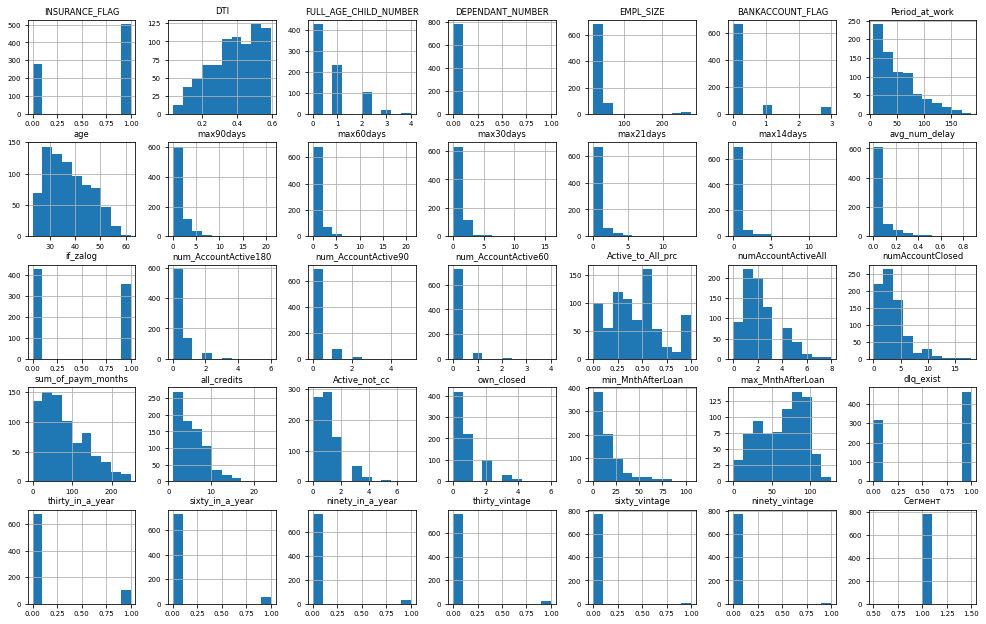


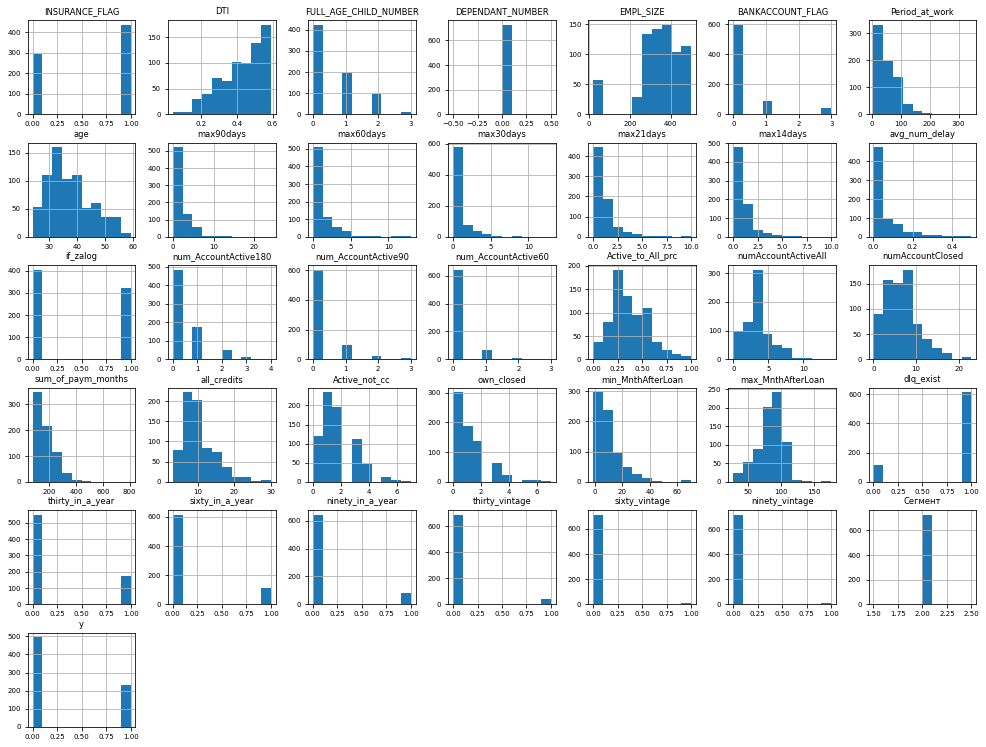
И отдельно гистограммы для числовых признаков по каждому из сегментов.

Сегмент 1:

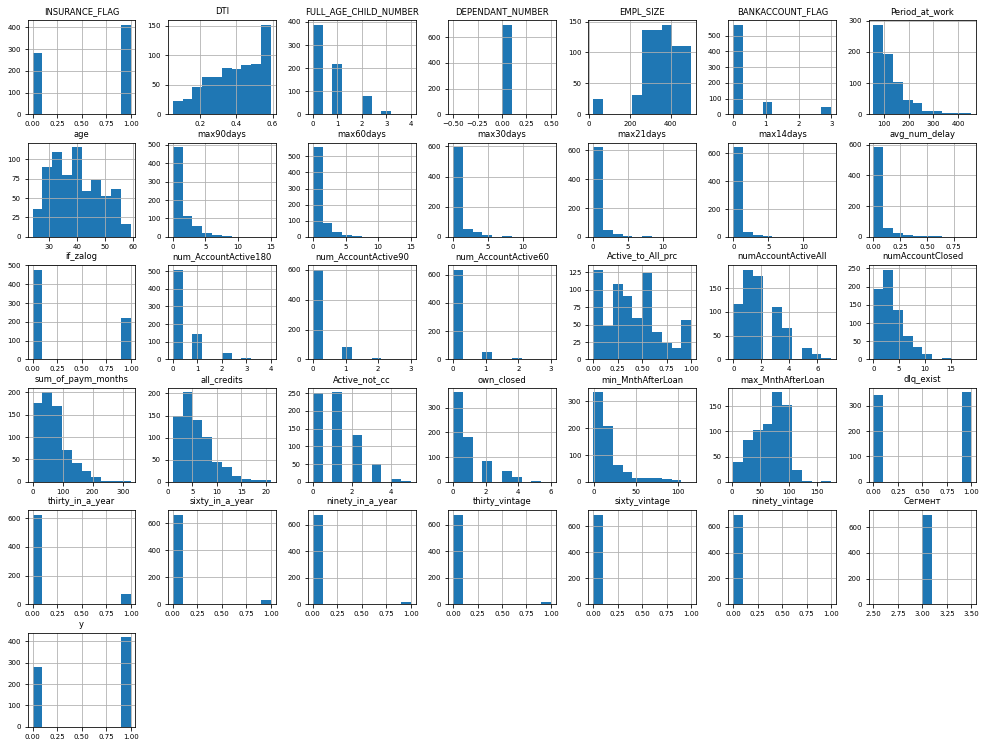


Сегмент 2

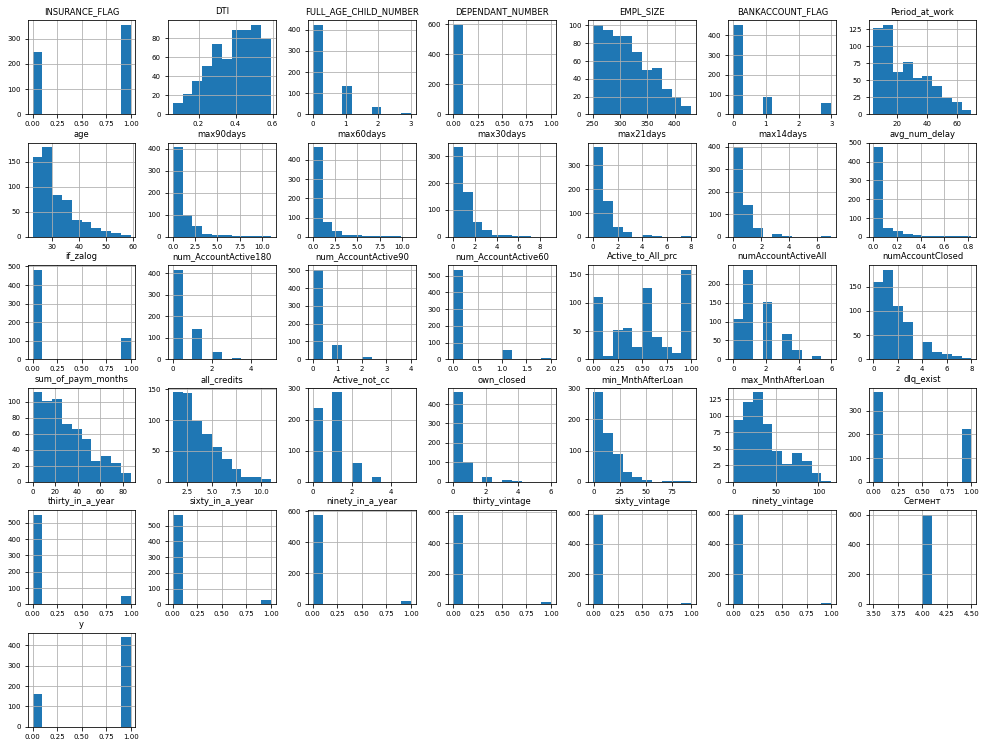


Сегмент 3

Сегмент 4



Сегмент 5



### **Описание сегментов**

Сегмент 1: В этом сегменте самая высокая доля замужних/женатых людей и клиентов со средне-специальным образованием, рабочих, наибольшее количество кредитных карт,

Сегмент 2: Чаще всех берут кредит под залог, наиболее давние клиенты банка, чаще всех оформляют страховки, преимущественно мужчины, в основном специалисты.

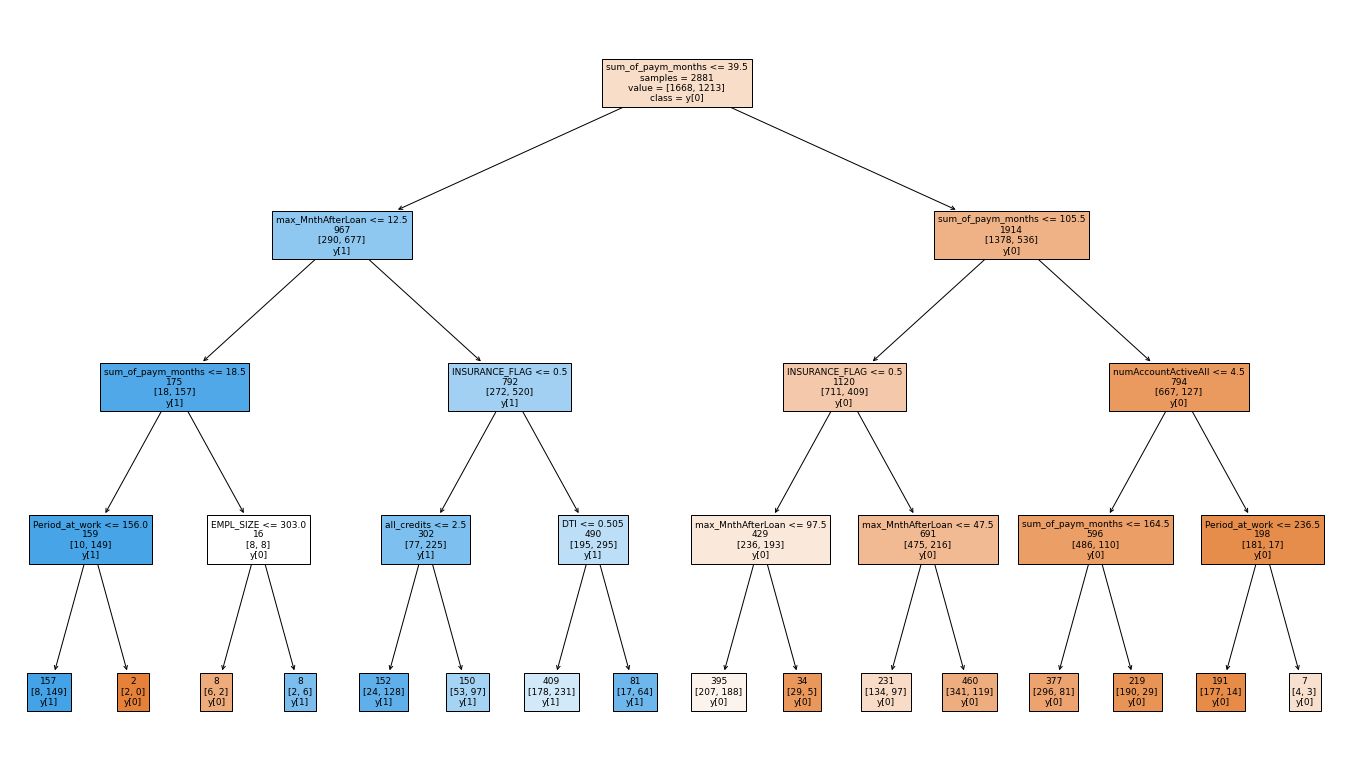
Сегмент 3: Наибольшее отношение долга к доходам, почти все имеют просрочку выплат по кредиту на данный момент, наибольшее количество кредитов. В этом сегменте примерно равное количество специалистов, менеджеров высшего и среднего звена

Сегмент 4: Наибольшее количество детей у клиентов, большая часть клиентов брали кредит недавно, примерно равное количество клиентов с просрочками выплат и без, наибольший стаж работы, примерно равное количество клиентов женского и мужского пола.

Сегмент 5: Наибольшее количество людей до 30, чаще всего не имеют залога, наименьшая сумма выплат по кредитам за месяц, преобладают клиенты женского пола

## **Решающее дерево**

Построим решающее дерево для определение сегментов клиентов, которые с наименьшей вероятностью будут просрочивать платежи.



Дерево выделило следующие сегменты клиентов, которые скорее всего будут вовремя вносить платежи по кредитам:

* Недавние (меньше года) клиенты банка, стаж работы меньше 13 лет, с суммой платежей по кредиту в месяц не больше 18 тысяч рублей.
* Давние клиенты банка со страховками и активной кредитной историей, суммой платежей не больше 39,5 тысяч рублей.
* Давние (больше года) клиенты, у которых не больше 2 кредитов, не берут страховки. Ежемесячная выплата по кредитам не больше 39,5 тысяч рублей.

## **Обоснование выбора методов**

Дерево решений

|  |  |
| --- | --- |
| Плюсы | Минусы |
| Легко интерпретируется в бизнес-правило | Разделяющая граница, которую строит дерево, не всегда проходит наиболее оптимальным способом |
| Быстро учится | Очень легко переобучается |
| Поддерживает числовые и категориальные признаки | Нестабильно, очень зависит от малейших изменений в данных |

K means

|  |  |
| --- | --- |
| Плюсы | Минусы |
| Прост в реализации | Необходимо подбирать количество кластеров |
| Легко масштабируется | Плохо справляется с задачей, когда элемент принадлежит сразу к нескольким кластерам в равной степени |
| Быстро обучается | Чувствителен к выбросам |