АНАЛИЗ ДАННЫХ

по страховым полисам ВЗР

Проблема:

Исследование данных по страховым полисам ВЗР

Цель:

- ✓ провести анализ отчетных данных страховой компании
- ✓ дать предложения по изменению системы страхования
- ✓ определить ключевые метрики
- ✓ сформировать общую витрину данных
- ✓ выделить аномальные данные в витрине

Исходные данные:

Представлены данные страховой компании по страхованиям путешествий –

- по контрактам
- по убыткам (выплатам)
- по клиентам компании

Способ реализации:

- формирование витрины данных в заказанном формате, в ценовых характеристиках по текущему курсу доллара США ЦБ РФ, проведение анализа данных, выделение аномалии
- проведение кластеризации клиентов лучшим способом
- оценка результата применения изменения способа определения цены страхования ВЗР с использованием метода кластеризации
- определение ключевые метрики и рекомендаций использования нового метода

Порядок реализации

1 Подготовка данных, формирование витрины данных

3711 общее количество клиентов для анализа

- Витрина данных сформирована по всему объему данных
- Все ценовые характеристики приведены к единому формату – доллару США по текущему курсу ЦБ РФ
- Данные хорошего качества, никаких значительных, особых дополнительных действий по заполнению базы данных не проводилось, все данные заполнены, дублирующих строк не обнаружено
- Данные отдельных столбцов не могут считаться аномальными, поэтому аномалия данных определена по комплексу числовых показателей с использованием метода LOF. Далее аномалии могут проанализированы на общую состоятельность данных

363 (10%) — количество условно аномальных данных в итоговом датасете

2 Кластеризация контрактов

- Кластеризация проведена тремя методами, сделан их сравнительный анализ
- Наилучшие результаты показала кластеризация контрактов иерархическим методом
- Оптимальное количество кластеров – 4
- Основные факторы кластеризации – продолжительность страховки, цена, страны путешествия
- На кластеризацию практически не повлияли пол, возраст, фактор наличия убытков

3 Применение кластеризации

- Применен новый метод формирования цены стоимости полиса ВЗР с использованием кластеризации

Порядок реализации

4 Анализ результатов применения нового подхода к формированию цены полиса ВЗР

1949 — количество клиентов с традиционным подходом к формированию цены полиса

1667 — количество клиентов с новым подходом к формированию цены полиса

Влияющие факторы:

- цена полиса
- конверсия в оформления
- убыточность

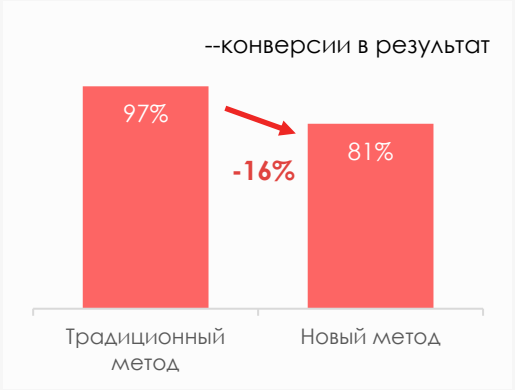
Оценка результатов применения нового подхода осуществлена с применением метода A/B- тестирования (проверки гипотез)

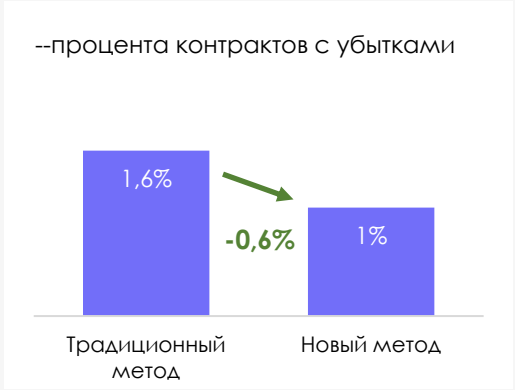
Выводы по A/B-тестированию:

1. Суммы убытков в обоих случаях равноценны при заданном уровне значимости
2. Количество случаев выплат отличаются по методикам
3. Цены и суммы страхования отличаются в обоих случаях
4. Конверсии в результат отличаются по методикам

Далее более подробно результаты применения нового метода в части конверсии в результат

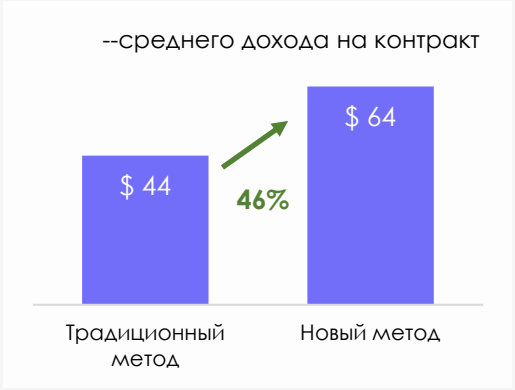
Анализ конверсии





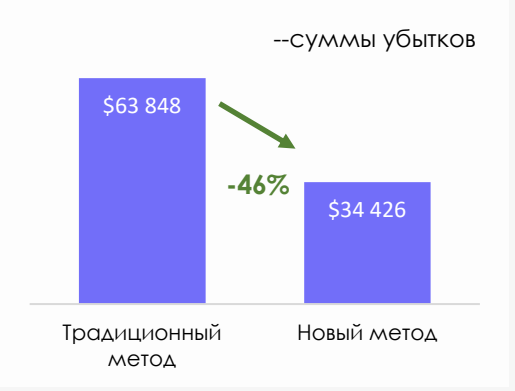
	Традиционный метод	Новый метод	Сравнение методов
Общее количество контрактов	1 949	1 667	
Количество действующих контрактов	1 882	1 338	
--- из них с убытками	30	14	
Процент конверсии в результат	**97%**	**81%**	**-17%**
Процент количества контрактов с убытками к действующим	**1,6%**	**1,%**	**-0,6%**

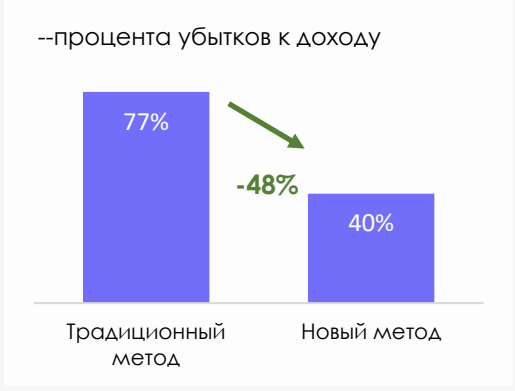
Анализ доходов



	Традиционный метод	Новый метод	Сравнение методов
Сумма дохода по действующим контрактам	82 911	86 267	
Средний доход на контракт	**44,1**	**64,5**	**46%**
Сумма по "потерянным" контрактам	**735**	**34 383**	**4578%**

Анализ убытков





	Традиционный метод	Новый метод	Сравнение методов
Сумма убытков по контрактам	63 848	34 426	**-46%**
Средняя сумма убытков на все контракты	34	26	**-24%**
Процент суммы убытков к сумме дохода	**77%**	**40%**	**-48%**

Заключение

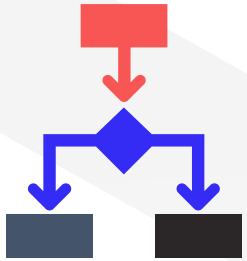
- Основной проблемой являлись большие суммы выплат по страховке — 77% от общей суммы доходов
- При этом количество страховых случаев было небольшим — всего 1,6%, а конверсия в страховку достигала 96%
- При изменении системы определения цены страхования и суммы страховки, процент конверсии существенно (но не критично) снизился — на 16% и достиг 80%, и потеря дохода по несостоявшимся контрактам увеличилась в 4 раза
- Но при этом удалось снизить количество страховых случаев (на 0,5%) и (что самое главное), существенно снизить суммы выплат по страховым случаям — почти в два раза — с 77% до 40%!
- Также вырос доход на один контракт — на 46%

Потеря дохода по несостоявшимся страховкам несоизмерима с размером выплат по страховым случаям, поэтому рекомендовано использование новую методику определения стоимости

Показатель, с которым следует работать в первую очередь, - это снижение размера страховых выплат, а потом уже конверсия.

Итоговый вывод и предложения по результатам исследования

- Методика кластеризации клиентов при формировании цены и размера страховки полностью подтверждает свою эффективность и продуктивность
- Следует продолжить работать в данном направлении и далее, снижая количество страховых выплат при достаточном уровне конверсии в результат и достижением требуемого уровня дохода



АЛГОРИТМ

выполнения работы,
подробности

1. Исследование представленных исходных датасетов, подготовка данных для аналитики

1.1. Формирование данных по курсам валют с сайта ЦБ РФ

- Сформирован путь приемки данных по курсам валют с сайта ЦБ РФ
- Определена функция и сформирован датасет курса валют

```
# API с курсами валют
r = requests.get('https://cbr.ru/scripts/XML_daily.asp?')
```

	rate_dt	currency_id	currency_cd	nominal_qty	currency_nm	currency_rate
0	06.10.2023	36	AUD	1	Австралийский доллар	63.5137
1	06.10.2023	944	AZN	1	Азербайджанский манат	58.6331
2	06.10.2023	826	GBP	1	Фунт стерлингов Соединенного королевства	120.7577

1.2. Объединение данных в единую витрину

- За основу взята таблица контрактов, которая была дополнена данными по клиентам, убыткам.
- Также таблица была дополнена данными по ценовым характеристикам (цена, сумма страховки, убытки), переведенным в доллары США по текущему курсу
- Из витрины были удалены «лишние» данные – наименование (оно было одно и тоже по всем данным), ценовые характеристики, неприведенные к единому формату, наименование валюты контракта.
- Следует также отметить, что количество id контрактов = количеству id клиентов, т.е. данные практически идентичны по сути.

1.3. Data cleaning датасета

Проведены мероприятия по проверке состоятельности данных ---

- База заполнена полностью, дублирующих строк нет
- Необоснованно пустых значений нет, пустоты заполнены нулем либо общими знаками в части ФИО
- Исследование аномалий ----
    - Исследованы все поля датасета на предмет аномальности.
    - Выделены выбросы по ценовым характеристикам, но они впрямую не могут быть признаны аномальными
    - Есть смысл сделать групповые исследования по числовым характеристикам
    - При оценке общими методами установлены данные, которые могут быть признаны аномальными.
    - По методу LOF количество таких данных - 363 строк (10%) и эта цифра похожа на возможную при оценке состава и качества данных датафрейма. По методу Изолирующего леса количество аномальных возросло - до 513 (14%), что вероятнее всего избыточно. В связи с эти оценка аномальности данных принята по методу LOF.
- В витрину данных добавлен признак аномальности данных по методу LOF

1.4. Формирование итоговой витрины данных

- Сформирована итоговая витрина данных для дальнейшей работы ---- 3711 строк, 16 столбцов

2. Кластеризация данных

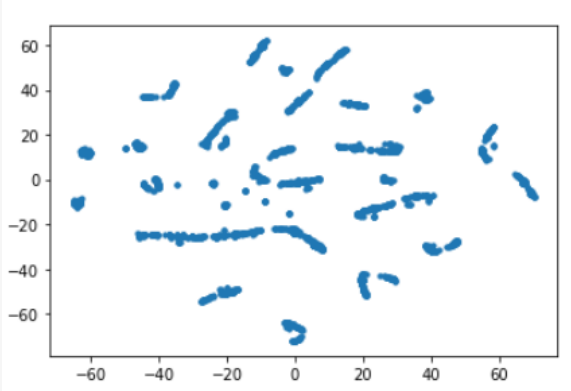
2.1. Обработка данных перед кластерным анализом

- Для кластеризации оставляем только основные столбцы – числовые и из качественных – признак страны, остальные столбцы удаляем, в т.ч. client_id, contract_id - столбцы идентичные id датасета
- Проводим замену признака пола на 0 - мужской и 1 - женский пол, а также замену статуса конверсии - 0- контракт завершен, 1-контракт действует
- Преобразовываем категориальный признак стран в числовой методом OneHotEncoder()
- В завершении подготовки данные стандартизируем StandardScaler()

2.2. Кластерный анализ

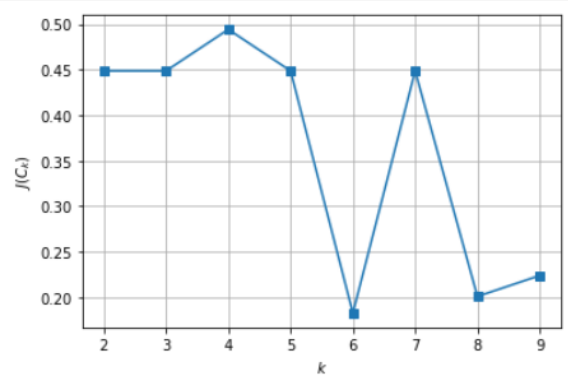
2.2.1. Снижаем размерность, применяя метод TSNE

- Получаем результат, который уже можно далее кластеризовать



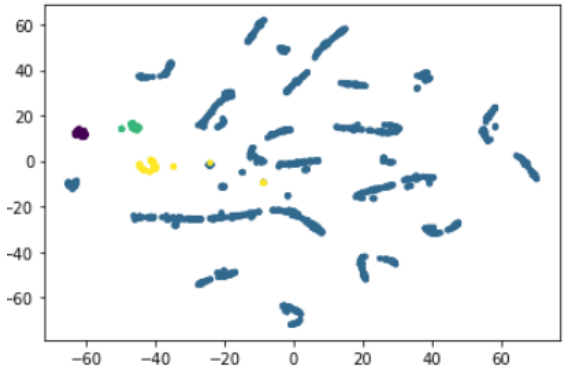
2.2.2. Определяем количество кластеров методом Локтя

- Принимаем оптимальное количество кластеров = 4



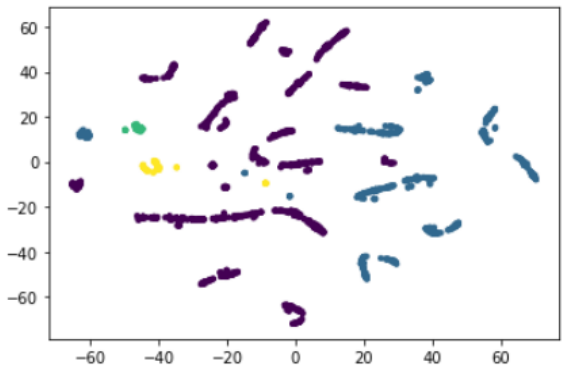
2.2.3. Кластеризация методом K-means

- Визуализация результата ----



2.2.4. Кластеризация иерархическим методом

- Визуализация результата ----

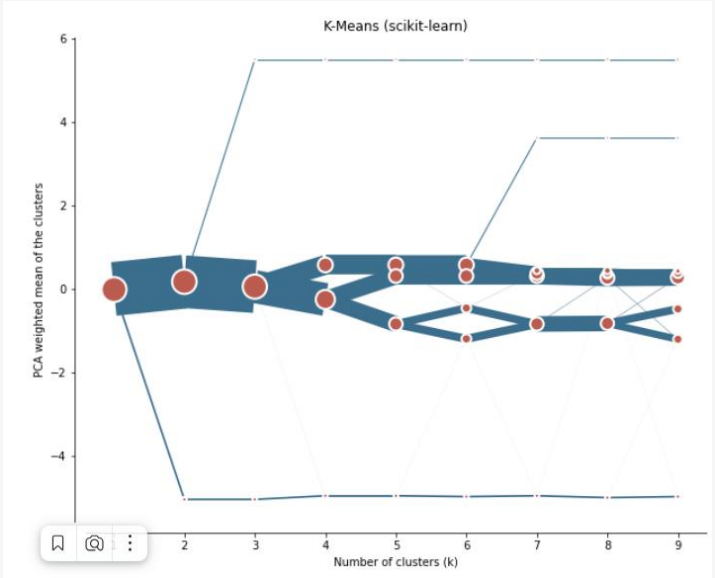


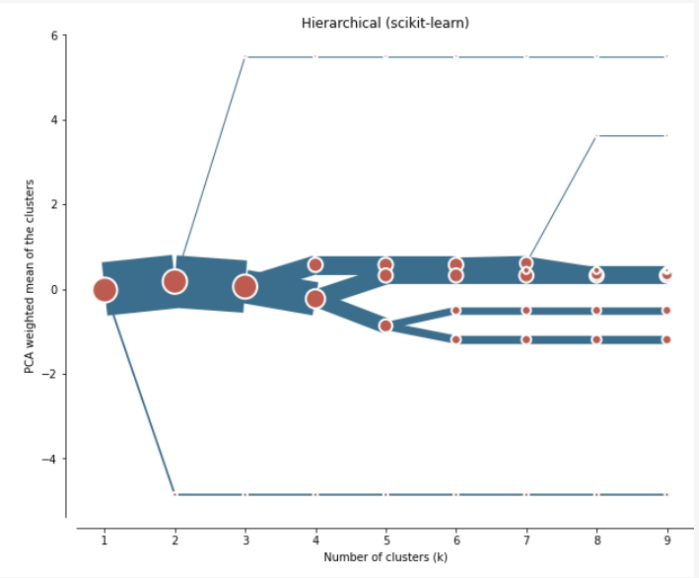
- Была также проведена кластеризация методом KPrototypes — результат аналогичен методу K-means
- Как видно из рисунков — результат кластеризации иерархическим методом выглядит более убедительно, но проверяем этот вывод отдельно

2.2.5. Сравнение методов кластеризации

- Сравнение результатов кластеризации через визуализацию методом Clustergram ---

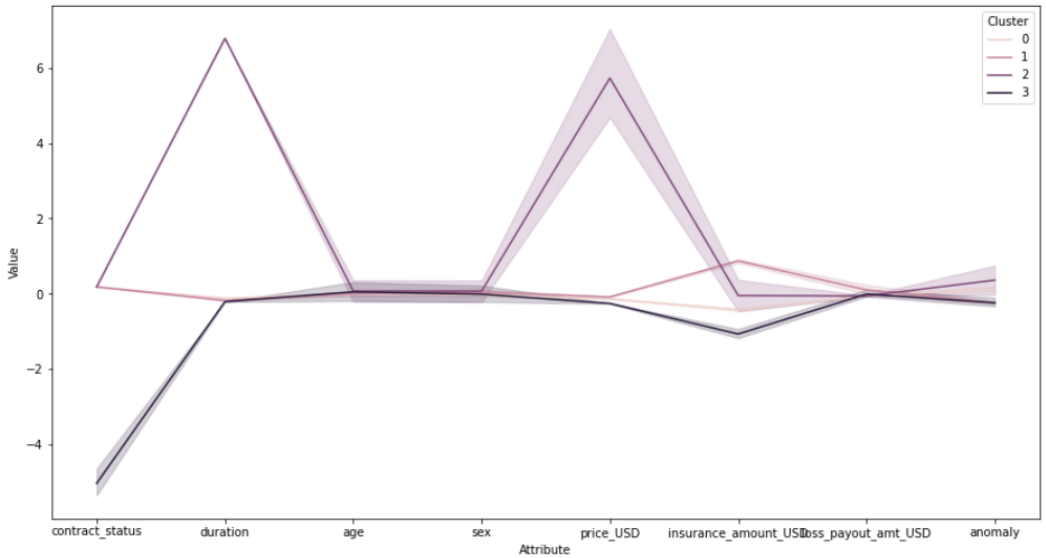
- Таким образом, подтвержден предварительный вывод – кластеризация иерархическим методом дает лучший результат с количеством кластеров = 4

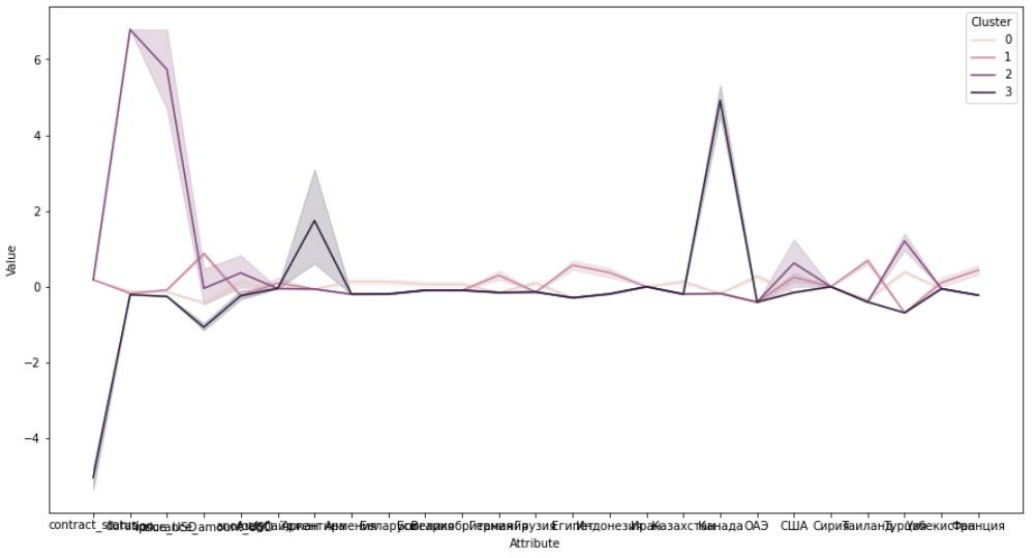




2.2.5. Визуализация итогов кластеризации

- Основными признаками кластеризации являются срок страхования, цена страховки и страна





3. A/B-тестирование старого и нового подходов к формированию стоимости полиса ВЗР

3.1. Подготовка данных

- в витрину добавлен признак группы, датасет разделен на два – control и test

3.2. Исследование гипотез

3.2.1. первая гипотеза по ценам страховки

- H0 - цены страхования равноценны как при традиционной методике, так и при использовании кластеризации с заданным уровнем оценки в 5%
- H1 - цены страхования отличаются при заданном уровне оценки в 5%

```
ttest_ind(control['price_USD'], test['price_USD'], equal_var= False)
```

```
TtestResult(statistic=-2.352614853341849, pvalue=0.018699878307990414, df=3326.506246589711)
```

- **Вывод**: т.к. p-value <0,05, мы не можем принять гипотезу о равенстве цен и суммы страховки в выборках с экспериментом и без него, т.е. цены страхования отличаются

3.2.2. вторая гипотеза по сумме страхования

- H0 - суммы страхования равноценны как при традиционной методике, так и при использовании кластеризации с заданным уровнем оценки в 5%
- H1 - суммы страхования отличаются при заданном уровне оценки в 5%

```
ttest_ind(control['insurance_amount_USD'], test['insurance_amount_USD'], equal_var= False)
```

```
TtestResult(statistic=11.69056492083666, pvalue=5.641803952473376e-31, df=3351.427757821664)
```

- **Вывод**: т.к. p-value <0,05, мы не можем принять гипотезу о равенстве суммы страховки в выборках с экспериментом и без него, т.е. суммы страхования отличаются

3.2.3. третья гипотеза по суммам убытков

- H0 - суммы убытков равноценны как при традиционной методике, так и при использовании кластеризации с заданным уровнем оценки в 5%
- H1 - суммы убытков отличаются при заданном уровне оценки в 5%

```
ttest_ind(control['loss_payout_amt_USD'], test['loss_payout_amt_USD'], equal_var= False)
```

```
TtestResult(statistic=0.7894536904578213, pvalue=0.4298993795144862, df=3565.529728050716)
```

- **Вывод:** т.к. p-value существенно >0,05, мы с большой долей вероятности можем предположить, что сумма убытков не зависит от проведения эксперимента (с большой долей вероятности принимаем верность нулевой гипотезы)

3.2.4. четвертая гипотеза о конверсии в результат

- H0 - конверсии в результат равноценны как при традиционной методике, так и при использовании кластеризации с заданным уровнем оценки в 5%
- H1 - конверсии в результат отличаются при заданном уровне оценки в 5%

```
ttest_ind(control['contract_status'], test['contract_status'], equal_var= False)
```

```
TtestResult(statistic=15.392006679043531, pvalue=6.554122494534892e-51, df=2254.721371927242)
```

- **Вывод:** т.к. p-value существенно <0,05, мы не можем принять гипотезу о равенстве конверсий в выборках с экспериментом и без него

3.2.5. Общие выводы по A/B тестированию

- Конверсия в результат выше в традиционном расчете (80% тест и 96,6% в контрольной выборке, разница в 16,3%, т.е. около 272 условно "потерянных" контрактов)
- Суммы убытков в обоих случаях с равноценны при заданном уровне значимости
- Цены и суммы страхования отличаются в обоих случаях

4. Общая аналитика по результатам

- Сравнительный анализ показателей явно демонстрирует эффективность нового подхода

```
ПОКАЗАТЕЛИ ПО ДАННЫМ  С ТРАДИЦИОННЫМ МЕТОДОМ
Количество контрактов  --- 1949
Количество контрактов с убытками  --- 30
Процент кол-ва контрактов с убытками к действ. контрактам  --- 1.5940488841657812 %
Сумма дохода по действующим контрактам --- 82911.26000000001 долл.
Сумма по потерянным контрактам  --- 734.9900000000001 долл.
Сумма убытков по контрактам  --- 63848.01000000001 долл.
Процент суммы убытков к сумме дохода по действ. контрактам --- 77.00764648854691 %
```

```
ПОКАЗАТЕЛИ ПО ДАННЫМ С КЛАСТЕРИЗАЦИЕЙ (НА ТЕСТЕ)
Количество контрактов  --- 1667
Количество контрактов с убытками  --- 14
Процент кол-ва контрактов с убытками к действ. контрактам  --- 1.046337817638266 %
Сумма дохода по действующим контрактам  --- 86266.65 долл.
Сумма по потерянным контрактам  --- 3438.13 долл.
Сумма убытков по контрактам --- 34425.91 долл.
Процент суммы убытков к сумме дохода по действ. контрактам --- 39.90639488145188 %
```