**Аналитический отчет**

Оглавление

[**1.** **Ситуация** 1](#_Toc147513491)

[**2.** **Проблема** 1](#_Toc147513492)

[**3.** **Порядок реализации** 2](#_Toc147513493)

[**3.1.** **Подготовка данных для аналитики** 2](#_Toc147513494)

[**3.1.1.** **Формирование данных по курсам валют с сайта ЦБ РФ** 2](#_Toc147513495)

[**3.1.2.** **Исследование представленных исходных датасетов** 2](#_Toc147513496)

[**3.1.3.** **Объединение данных в единую витрину** 3](#_Toc147513497)

[**3.1.4.** **Data cleaning датасета** 3](#_Toc147513498)

[**3.1.5.** **Формирование итоговой витрины данных** 3](#_Toc147513499)

[**3.2.** **Кластеризация данных** 3](#_Toc147513500)

[**3.2.1.** **Обработка данных перед кластерным анализом** 3](#_Toc147513501)

[**3.2.2.** **Кластерный анализ** 4](#_Toc147513502)

[**3.3.1.** **Подготовка данных** 7](#_Toc147513503)

[**3.3.2.** **Исследование гипотез** 7](#_Toc147513504)

[**3.3.3.** **Общая аналитика по результатам анализа** 8](#_Toc147513505)

[**4.** **Заключение** 8](#_Toc147513506)

[**4.1.** **Общие результаты исследования -** 8](#_Toc147513507)

[**4.2.** **Результирующий вывод -** 9](#_Toc147513508)

# **Ситуация**

* 1. Для анализа были представлены данные выгрузки отчетных данных страховой компании, определяющих характеристики страхования путешествий, базу информации по покупателям, а также страховым выплатам (убыткам).

Были представлены следующие файлы

- CASE\_CONTRACTS.xlsx – информация по контрактам

- CASE\_LOSSES.xlsx - данные по убыткам

- CASE\_CLIENTS.xlsx – информация по клиентам компании

1.2. Также был представлен файл (experiment\_group.xlsx) - разделение контрактов на группы, которое было сделано при применении методики изменения цены и суммы страхования после кластеризации клиентов

# **Проблема**

Необходимо

- сформировать общую витрину данных из набора представленных файлов

- учесть при формировании витрины возможность перевода данных в доллары США по текущему курсу валюты ЦБ РФ

- провести полный анализ исходных данных, оценку, клининг, заполнение

- определить лучшую методику кластеризации данных

- провести тестирование старого и нового подходов к формированию стоимости полиса

- сделать итоговые выводы и дать рекомендации по применению нового подхода к формированию стоимости полиса, а также основным метрикам оценки результата

# **Порядок реализации**

## **Подготовка данных для аналитики**

### **Формирование данных по курсам валют с сайта ЦБ РФ**

1. Сформирован путь приемки данных по курсам валют с сайта ЦБ РФ



2. Определена функция и сформирован датасет курса валют

Изображение выглядит как текст, Шрифт, число, линия

Автоматически созданное описание

### **Исследование представленных исходных датасетов**

#### 3.1.2.1. Таблица контрактов

Таблица контрактов – 3 711 контрактов, 10 столбцов данных

Изображение выглядит как текст, Шрифт, линия, число

Автоматически созданное описание

#### 3.1.2.2. Таблица клиентов

Таблица клиентов – 3 711 клиентов, 6 столбцов данных

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

#### 3.1.1.3. Таблица убытков

Таблица убытков – 45 строк данных, 5 столбцов данных

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

#### 3.1.1.4. Таблица разделения данных на группы

Таблица групп – 14 842 строки данных, 4 столбца данных

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

### **Объединение данных в единую витрину**

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

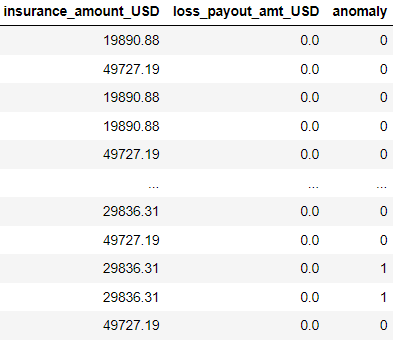
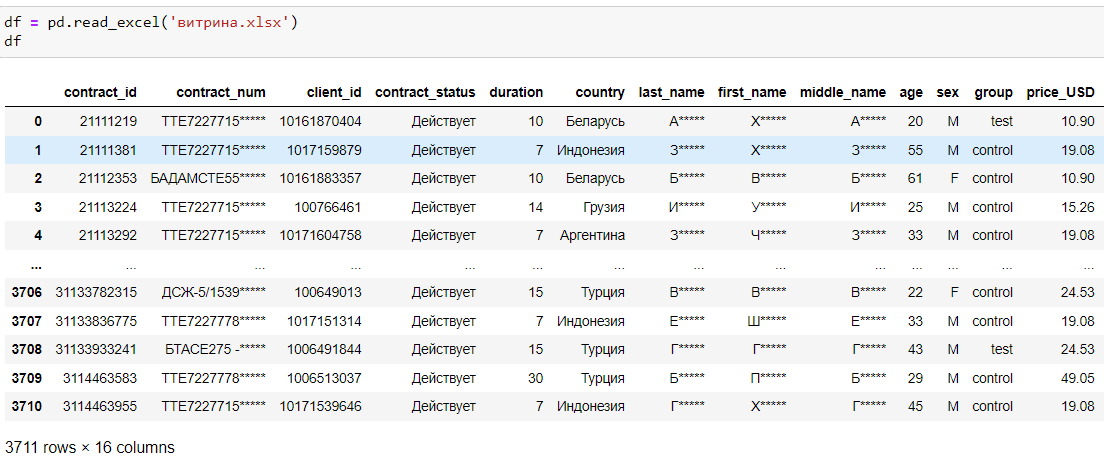
### **Data cleaning датасета**

Проведены мероприятия по проверке состоятельности данных ---

1. База заполнена полностью, дублирующих строк нет
2. Необоснованно пустых значений нет, пустоты заполнены нулем либо общими знаками в части ФИО
3. Исследование аномалий ----
   1. Исследованы все поля датасета на предмет аномальности.
   2. Выделены выбросы по ценовым характеристикам, но они впрямую не могут быть признаны аномальными
   3. Есть смысл сделать групповые исследования по числовым характеристикам
   4. При оценке общими методами установлены данные, которые могут быть признаны аномальными.
   5. 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.
4. В витрину данных добавлен признак аномальности данных по методу LOF
5. В витрину добавлен признак группы, т.к. кластеризацию проводим по группе контрактов с традиционным методом формирования цены полиса ВЗР

### **Формирование итоговой витрины данных**

Сформирована итоговая витрина данных для дальнейшей работы



## **Кластеризация данных**

### **Обработка данных перед кластерным анализом**

* Кластеризацию проводим в нетестовой группе
* 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()

### **Кластерный анализ**

* + - 1. **Снижаем размерность, применяя метод TSNE.**

Изображение выглядит как снимок экрана, линия, диаграмма, График

Автоматически созданное описаниеПолучаем результат, который уже можно далее кластеризовать –

Изображение выглядит как линия, График, диаграмма, число

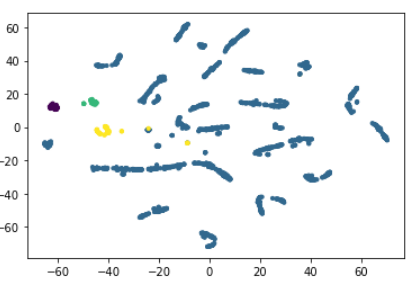
Автоматически созданное описание

* + - 1. **Определяем количество кластеров методом Локтя.**

Принимаем оптимальное количество кластеров = 4

**3.2.2.3. Кластеризация методом К-means**

Результат кластеризации приведен на рисунке ---



* + - 1. **Кластеризация иерархическим методом**

Результат кластеризации приведен на рисунке --

Изображение выглядит как снимок экрана, линия, диаграмма, Красочность

Автоматически созданное описание

Была также проведена кластеризация методом KPrototypes – результат аналогичен методу К-means.

Как видно из рисунков – результат кластеризации иерархическим методом выглядит более убедительно, но проверим этот вывод отдельно.

* + - 1. **Сравнение методов кластеризации**

Сравним результаты кластеризации через визуализацию методом Clustergram ---

Изображение выглядит как линия, снимок экрана, текст, диаграмма

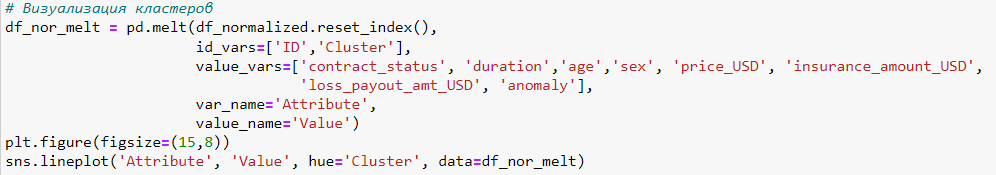
Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Таким образом, подтвержден предварительный вывод – кластеризация иерархическим методом дает лучший результат.

* + - 1. **Визуализация результатов иерархической кластеризации**

- Так как признаков стран много, проведем визуализацию без них -



Изображение выглядит как диаграмма, линия, оригами

Автоматически созданное описание

##### Как видно из графика - кластеры практически не отличаются по полу и возрасту, а также по величине убытков, поэтому уберем эти поля из рассмотрения и добавим страны.

##### Также из графика явно видно, что основные признаки кластеризации - это срок страхования и цена страховки

- Уберем совпадающие признаки и добавим страны в визуализацию - -

Изображение выглядит как диаграмма, График, линия

Автоматически созданное описание

Из данного графика видно, что кластеризация непосредственно проведена и по признаку стран.

* + - 1. **Выводы по итогам кластеризации**
* Таким образом, как видно из кластерграмм иерархический метод показывает явно более качественные результаты, чем метод К-средних.
* Кластеризацию стоит делать именно по результатам иерархического метода с количеством кластеров =5, увеличивать количество кластеров нецелесообразно
* Основными признаками кластеризации являются срок страхования, цена страховки и страна

**3.3. А/В-тестирование старого и нового подходов к формированию стоимости полиса ВЗР**

Первый подход включает в себя традиционную оценку рисков, а второй — использование результатов кластеризации.

Основные влияющие факторы: цена полиса, конверсия в оформления и убыточность.

### **Подготовка данных**

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

### **Исследование гипотез**

* + - 1. **Исследуем первую гипотезу по ценам страховки**

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

Н1 - цены страхования отличаются при заданном уровне оценки в 5%

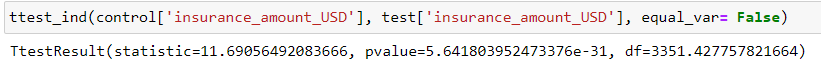


#### Вывод: При сравнении выборок по цене страховки, т.к. p-value <0,05, мы не можем принять гипотезу о равенстве цен и суммы страховки в выборкахс экспериментом и без него, т.е. цены страхования отличаются

* + - 1. **Исследуем вторую гипотезу по сумме страхования**

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

Н1 - суммы страхования отличаются при заданном уровне оценки в 5%

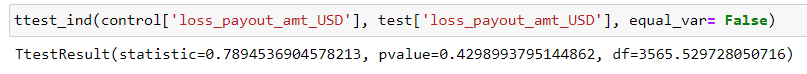


**Вывод:** При сравнении выборок по общей сумме страховки, т.к. p-value <0,05, мы не можем принять гипотезу о равенстве суммы страховки в выборках с экспериментом и без него, т.е. суммы страхования отличаются

* + - 1. **Исследуем третью гипотезу по суммам убытков**

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

Н1 - суммы убытков отличаются при заданном уровне оценки в 5%

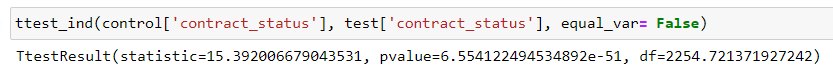


**Вывод:** При сравнении выборок по стоимости убытков, т.к. p-value существенно >0,05, мы с большой долей вероятности можем предположить, что сумма убытков не зависит от проведения эксперимента (с большой долей вероятности принимаем верность нулевой гипотезы)

* + - 1. **Исследуем четвертую гипотезу о конверсии в результат**

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

Н1 - конверсии в результат отличаются при заданном уровне оценки в 5%



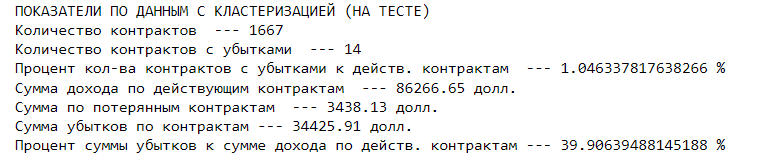
**Вывод:** При сравнении выборок по конверсии в продажи, т.к. p-value существенно <0,05, мы не можем принять гипотезу о равенстве конверсий в выборках с экспериментом и без него.

* + - 1. **Общие выводы по A/B тестированию**

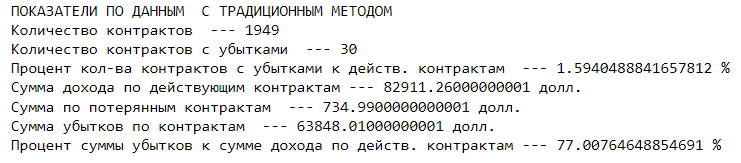
1. Конверсия в результат выше в традиционном расчете (80% тест и 96,6% в контрольной выборке, разница в 16,3%, т.е. около 272 условно "потерянных" контрактов)
2. Суммы убытков в обоих случаях с равноценны при заданном уровне значимости
3. Цены и суммы страхования отличаются в обоих случаях

### **Общая аналитика по результатам**

* + - 1. **Показатели данных выборки с применением нового подхода к формированию стоимости**



* + - 1. **Показатели данных выборки с применением старого подхода к формированию стоимости**



Сравнительный анализ показателей явно демонстрирует эффективность нового подхода

# **Заключение**

## **Общие результаты исследования -**

* 1. Основной проблемой являлись большие суммы выплат по страховке - 77% от общей суммы доходов
  2. При этом количество страховых случаев было небольшим - всего 1,6%, а конверсия в страховку достигала 96%
  3. При изменении системы определения цены страхования и суммы страховки, процент конверсии существенно (но не критично) снизился - на 16% и достиг 80%, а потеря дохода по несостоявшимся контрактам увеличилась в 4 раза
  4. 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%!
  5. Потеря дохода по несостоявшимся страховкам несоизмерима с размером выплат по страховым случаям, поэтому следует использовать новую методику определения стоимости
  6. Также, показатель, с которым следует работать в первую очередь, - это снижение размера страховых выплат, а потом уже конверсия.

## **Результирующий вывод -**

#### Таким образом, 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.