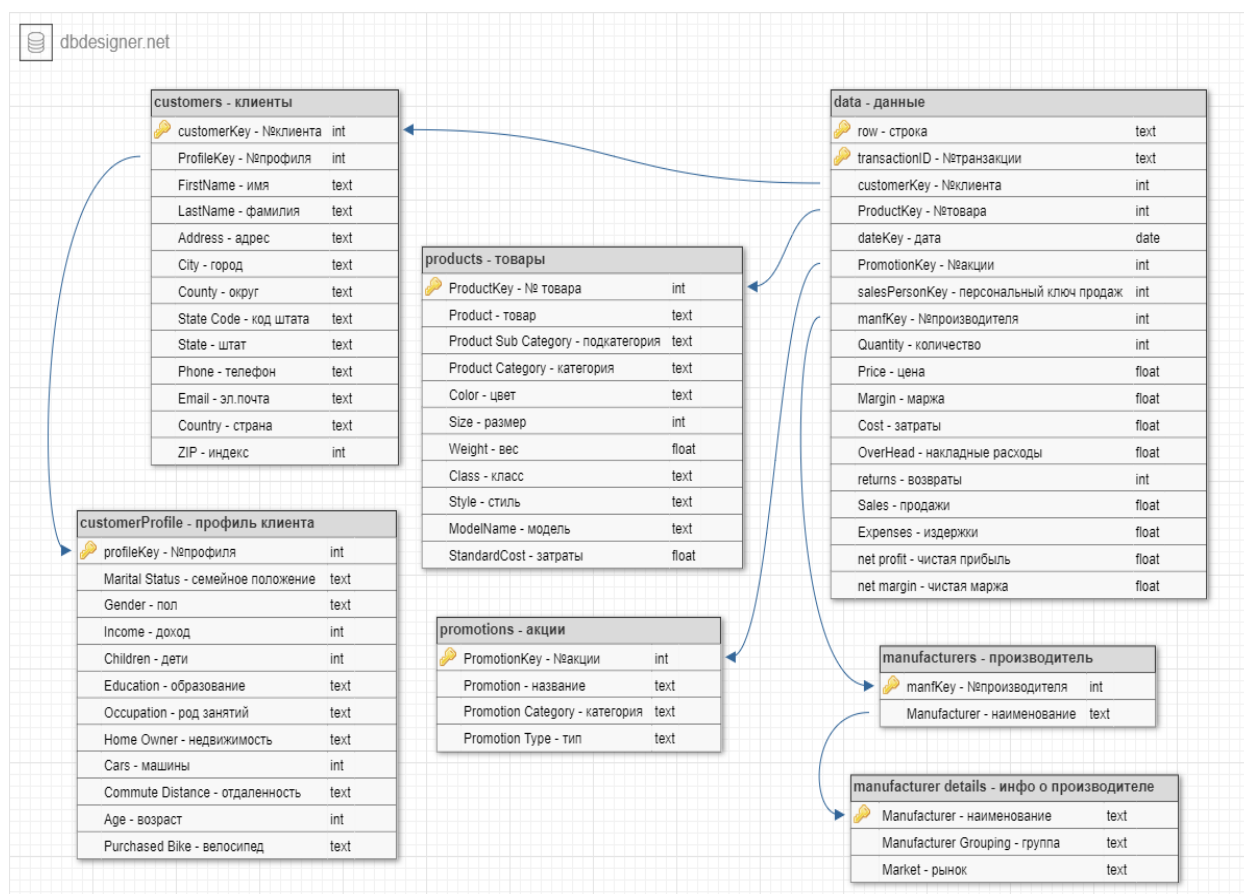


1. Описание кейса и постановка задач

Для работы мной был выбран кейс №5. В нем представлена информация о продажах магазина, который реализует велосипеды и все, что связано с ездой на велосипедах: экипировка, запчасти, аксессуары. Выборка данных хранит информацию о продажах за период с 2008 по 2010 годы.

Чтобы лучше понять кейс и взаимосвязи между таблицами построим схему базы данных. Для этого используем бесплатный сервис dbdesigner.net (<https://app.dbdesigner.net/>).



Сформулируем задачи для анализа:

- 1) Спрогнозировать продажи на следующий квартал.
- 2) Построить модель, которая будет предсказывать вероятность возврата проданного товара.
- 3) Провести анализ клиентов, составить список приоритетных клиентов для формирования персональных предложений. Можно ли составить портрет «идеального» клиента?
- 4) Провести анализ ассортимента: по каким товарам необходимо сформировать запас ввиду их высокого спроса, а какие товары необходимо перевести на систему пред заказа, ввиду их низкой популярности.
- 5) Создать дашборд, отражающий продажи по странам и по периодам. Есть ли регионы, которые целесообразно покинуть?

2. EDA на платформе Loginom

Данные кейса хранятся на 7 листах. В первую очередь произведем загрузку и слияние всех таблиц. В частности, рассмотрим количество уникальных значений, временные периоды, пропуски и т.д.

Сразу бросается в глаза, что частично отсутствуют данные по столбцу Returns (это единственный столбец с пропусками). Так как анализ данного признака – это одна из наших задач, необходимо перед началом EDA заполнить данные пропуски (значением 0). Для этого используем блок «Заполнение пропусков». Далее с помощью раздела «Статистика» рассмотрим структуру данных о возвратах:

№	Метка	Доля	Кол-во	%
1	<null>		0	0
2	0		8886	89
3	1		792	8
4	2		322	3
5	<null>		0	0

Так как возвратов в количестве 2 штуки всего 3%, предлагаю сделать данный признак бинарным:

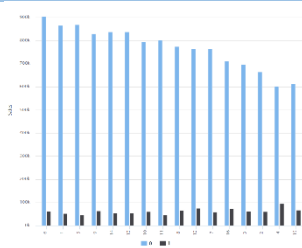
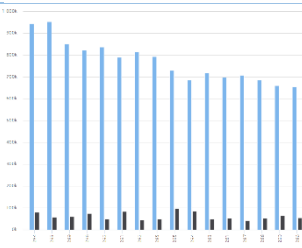
0 – если возврата не было

1 – если возврат был.

Для этого с помощью блока «Замена»: заменим значение 2 на 1.

Далее продолжим изучать данные с помощью возможностей блока «Визуализаторы» Loginom.

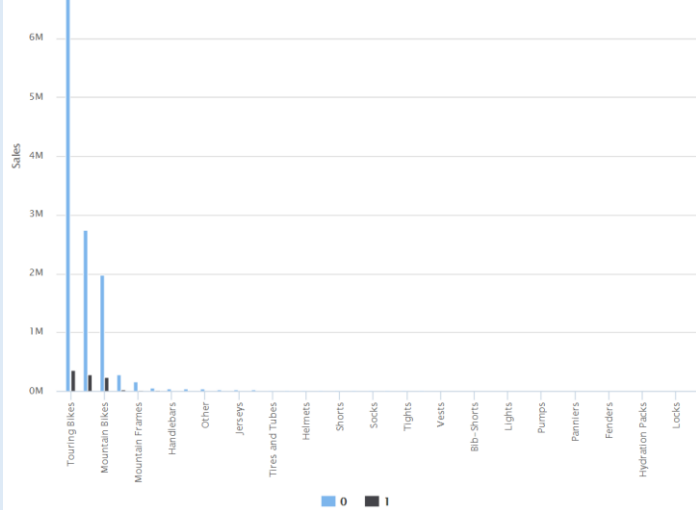
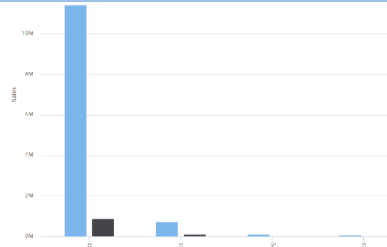
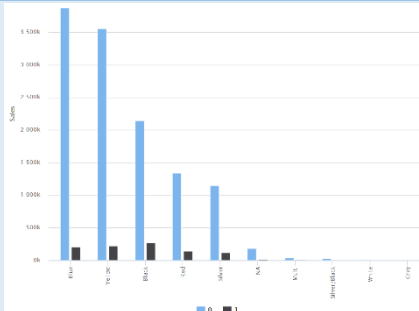
Лист 1 – data – исходный вариант содержал 18 столбцов и 10 000 строк. Это основной информационный лист с данными о продажах. Все остальные листы дополняют информацию о признаках листа data. Рассмотрим информацию подробнее:

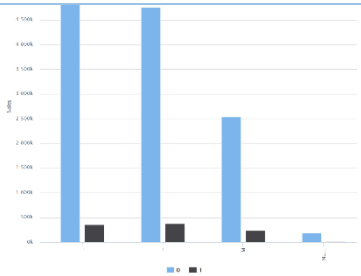
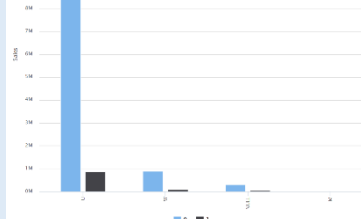
Наименование	Планируемые действия																														
Row Строка	Содержит 10 000 уникальных значений. Пригодится для RFM-анализа, поэтому оставим.																														
transactionID №транзакции	Содержит 9 955 уникальных значений. Только 45 из 10 000 проданных товаров были не единственными в чеке. Анализ ассоциативных правил будет не информативен. Удалим столбец.																														
customerKey №клиента	Цифровой идентификатор клиента. Моделью будет рассматриваться как число, а не как категориальный признак. Поэтому данный признак будет заменен на имя и фамилию из таблицы Customer.																														
ProductKey №товара	Цифровой идентификатор товара. Моделью будет рассматриваться как число, а не как категориальный признак. Поэтому данный признак будет заменен на наименование товара из таблицы Products.																														
dateKey Дата продажи	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>01.01.2008...</td><td></td><td>2104</td><td>21</td></tr><tr><td>3</td><td>31.12.2008...</td><td></td><td>4011</td><td>40</td></tr><tr><td>4</td><td>31.12.2009...</td><td></td><td>3885</td><td>39</td></tr><tr><td>5</td><td><null></td><td></td><td>0</td><td>0</td></tr></table> <p>Создав 36 интервалов, при рассмотрении гистограммы по данному признаку, видим, что в таблице данные за все месяца 2008 г., 2009 г., 2010 г. Пропусков нет. В 2009 г. продажи растут примерно в 2 раза к 2008 г. В 2010 - на уровне 2009г.</p>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	01.01.2008...		2104	21	3	31.12.2008...		4011	40	4	31.12.2009...		3885	39	5	<null>		0	0
№	Метка	Доля	Кол-во	%																											
1	<null>		0	0																											
2	01.01.2008...		2104	21																											
3	31.12.2008...		4011	40																											
4	31.12.2009...		3885	39																											
5	<null>		0	0																											
PromotionKey №акции	 <p>По данному показателю прослеживается взаимосвязь с продажами. Но сильной взаимосвязи с количеством возвратов нет, предлагаю удалить данный признак и соответственно все взаимосвязанные с ним данные.</p>																														
salesPersonKey Персональный ключ продаж	 <p>Данный критерий не опознан. Нет дополнительной таблицы, которая бы расшифровывала эти данные. В столбце представлены значения от 281 до 296. Но Сильной взаимосвязи с количеством возвратов нет, предлагаю удалить данный признак.</p>																														

manfKey №производителя		Цифровой идентификатор производителя. Моделью будет рассматриваться как число, а не как категориальный признак. Поэтому данный признак будет заменен на наименование производителя из таблицы manufacturers. Явно видно, что по некоторым производителям возвраты случаются чаще.																																																																																										
Quantity Количество	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>1</td><td></td><td>8803</td><td>88</td></tr><tr><td>3</td><td>2</td><td></td><td>350</td><td>4</td></tr><tr><td>4</td><td>3</td><td></td><td>331</td><td>3</td></tr><tr><td>5</td><td>4</td><td></td><td>301</td><td>3</td></tr><tr><td>6</td><td>5</td><td></td><td>63</td><td>1</td></tr><tr><td>7</td><td>6</td><td></td><td>53</td><td>1</td></tr><tr><td>8</td><td>7</td><td></td><td>51</td><td>1</td></tr><tr><td>9</td><td>8</td><td></td><td>48</td><td>0</td></tr><tr><td>10</td><td><null></td><td></td><td>0</td><td>0</td></tr></table>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	1		8803	88	3	2		350	4	4	3		331	3	5	4		301	3	6	5		63	1	7	6		53	1	8	7		51	1	9	8		48	0	10	<null>		0	0	В 88% случаев приобретается 1 единица товара.																																			
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	1		8803	88																																																																																								
3	2		350	4																																																																																								
4	3		331	3																																																																																								
5	4		301	3																																																																																								
6	5		63	1																																																																																								
7	6		53	1																																																																																								
8	7		51	1																																																																																								
9	8		48	0																																																																																								
10	<null>		0	0																																																																																								
Price Цена	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>5:353</td><td></td><td>6391</td><td>64</td></tr><tr><td>3</td><td>353:700</td><td></td><td>36</td><td>0</td></tr><tr><td>4</td><td>702:1050</td><td></td><td>0</td><td>0</td></tr><tr><td>5</td><td>1050:1299</td><td></td><td>593</td><td>6</td></tr><tr><td>6</td><td>1399:1747</td><td></td><td>1450</td><td>14</td></tr><tr><td>7</td><td>1747:2096</td><td></td><td>1026</td><td>10</td></tr><tr><td>8</td><td>2096:2444</td><td></td><td>108</td><td>1</td></tr><tr><td>9</td><td>2444:2793</td><td></td><td>75</td><td>1</td></tr><tr><td>10</td><td>2793:3141</td><td></td><td>53</td><td>1</td></tr><tr><td>11</td><td>3141:3489</td><td></td><td>26</td><td>0</td></tr><tr><td>12</td><td>3489:3838</td><td></td><td>8</td><td>0</td></tr><tr><td>13</td><td>3838:4186</td><td></td><td>128</td><td>1</td></tr><tr><td>14</td><td>4186:4535</td><td></td><td>85</td><td>1</td></tr><tr><td>15</td><td>4535:4883</td><td></td><td>13</td><td>0</td></tr><tr><td>16</td><td>4883:5232</td><td></td><td>48</td><td>0</td></tr><tr><td>17</td><td><null></td><td></td><td>0</td><td>0</td></tr></table> 	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	5:353		6391	64	3	353:700		36	0	4	702:1050		0	0	5	1050:1299		593	6	6	1399:1747		1450	14	7	1747:2096		1026	10	8	2096:2444		108	1	9	2444:2793		75	1	10	2793:3141		53	1	11	3141:3489		26	0	12	3489:3838		8	0	13	3838:4186		128	1	14	4186:4535		85	1	15	4535:4883		13	0	16	4883:5232		48	0	17	<null>		0	0	Более половины товаров (64%) продается по цене от 5 до 353 ден.ед. – это комплектующие, одежда и аксессуары. Примерно 30% товаров продаются по цене в диапазоне от 1000 до 2000 – это велосипеды.
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	5:353		6391	64																																																																																								
3	353:700		36	0																																																																																								
4	702:1050		0	0																																																																																								
5	1050:1299		593	6																																																																																								
6	1399:1747		1450	14																																																																																								
7	1747:2096		1026	10																																																																																								
8	2096:2444		108	1																																																																																								
9	2444:2793		75	1																																																																																								
10	2793:3141		53	1																																																																																								
11	3141:3489		26	0																																																																																								
12	3489:3838		8	0																																																																																								
13	3838:4186		128	1																																																																																								
14	4186:4535		85	1																																																																																								
15	4535:4883		13	0																																																																																								
16	4883:5232		48	0																																																																																								
17	<null>		0	0																																																																																								
Margin Плановая Маржа - наценка	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>0,330: 0,...</td><td></td><td>17</td><td>0</td></tr><tr><td>3</td><td>0,330: 0,...</td><td></td><td>21</td><td>0</td></tr><tr><td>4</td><td>0,348: 0,644</td><td></td><td>85</td><td>1</td></tr><tr><td>5</td><td>0,644: 0,925</td><td></td><td>2777</td><td>28</td></tr><tr><td>6</td><td>0,335: 0,426</td><td></td><td>4555</td><td>46</td></tr><tr><td>7</td><td>0,420: 0,618</td><td></td><td>1342</td><td>13</td></tr><tr><td>8</td><td>0,616: 0,809</td><td></td><td>36</td><td>1</td></tr><tr><td>9</td><td>0,809: 1,000</td><td></td><td>1106</td><td>11</td></tr><tr><td>10</td><td><null></td><td></td><td>0</td><td>0</td></tr></table>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	0,330: 0,...		17	0	3	0,330: 0,...		21	0	4	0,348: 0,644		85	1	5	0,644: 0,925		2777	28	6	0,335: 0,426		4555	46	7	0,420: 0,618		1342	13	8	0,616: 0,809		36	1	9	0,809: 1,000		1106	11	10	<null>		0	0	Чаще всего маржа составляет от 5 до 40 %.. В 11% случаев маржа составила 80-100%. Для анализа интересна только с точки зрения сравнения с реальной маржой.																																			
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	0,330: 0,...		17	0																																																																																								
3	0,330: 0,...		21	0																																																																																								
4	0,348: 0,644		85	1																																																																																								
5	0,644: 0,925		2777	28																																																																																								
6	0,335: 0,426		4555	46																																																																																								
7	0,420: 0,618		1342	13																																																																																								
8	0,616: 0,809		36	1																																																																																								
9	0,809: 1,000		1106	11																																																																																								
10	<null>		0	0																																																																																								
Cost Закупочная стоимость единицы товара	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>3:482</td><td></td><td>6125</td><td>61</td></tr><tr><td>3</td><td>482:961</td><td></td><td>271</td><td>3</td></tr><tr><td>4</td><td>961:1910</td><td></td><td>2094</td><td>21</td></tr><tr><td>5</td><td>1910:3819</td><td></td><td>710</td><td>8</td></tr><tr><td>6</td><td>3819:7638</td><td></td><td>123</td><td>1</td></tr><tr><td>7</td><td>7638:15277</td><td></td><td>118</td><td>1</td></tr><tr><td>8</td><td>15277:30554</td><td></td><td>30</td><td>1</td></tr><tr><td>9</td><td>30554:61108</td><td></td><td>101</td><td>1</td></tr><tr><td>10</td><td>61108:122216</td><td></td><td>23</td><td>0</td></tr><tr><td>11</td><td>122216:244432</td><td></td><td>5</td><td>0</td></tr><tr><td>12</td><td>244432:488864</td><td></td><td>8</td><td>0</td></tr><tr><td>13</td><td>488864:977728</td><td></td><td>2</td><td>0</td></tr><tr><td>14</td><td><null></td><td></td><td>0</td><td>0</td></tr></table>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	3:482		6125	61	3	482:961		271	3	4	961:1910		2094	21	5	1910:3819		710	8	6	3819:7638		123	1	7	7638:15277		118	1	8	15277:30554		30	1	9	30554:61108		101	1	10	61108:122216		23	0	11	122216:244432		5	0	12	244432:488864		8	0	13	488864:977728		2	0	14	<null>		0	0	В 64% случаев затраты на покупку товара составляю до 482 ден.ед. - это комплектующие, одежда и аксессуары. В 29% случаев от 960 до 2000 – это закупка велосипедов.															
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	3:482		6125	61																																																																																								
3	482:961		271	3																																																																																								
4	961:1910		2094	21																																																																																								
5	1910:3819		710	8																																																																																								
6	3819:7638		123	1																																																																																								
7	7638:15277		118	1																																																																																								
8	15277:30554		30	1																																																																																								
9	30554:61108		101	1																																																																																								
10	61108:122216		23	0																																																																																								
11	122216:244432		5	0																																																																																								
12	244432:488864		8	0																																																																																								
13	488864:977728		2	0																																																																																								
14	<null>		0	0																																																																																								
OverHead Накладные расходы	<table><tr><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>-1,3: 15,0</td><td></td><td>5116</td><td>51</td></tr><tr><td>15,0: 39,3</td><td></td><td>1530</td><td>15</td></tr><tr><td>39,3: 59,6</td><td></td><td>748</td><td>7</td></tr><tr><td>59,6: 79,8</td><td></td><td>533</td><td>5</td></tr><tr><td>79,8: 100,1</td><td></td><td>550</td><td>6</td></tr><tr><td>100,1: 120,4</td><td></td><td>420</td><td>4</td></tr><tr><td>120,4: 140,7</td><td></td><td>318</td><td>3</td></tr><tr><td>140,7: 160,9</td><td></td><td>202</td><td>2</td></tr><tr><td>160,9: 181,2</td><td></td><td>141</td><td>1</td></tr><tr><td>181,2: 201,5</td><td></td><td>110</td><td>1</td></tr><tr><td>201,5: 221,8</td><td></td><td>66</td><td>1</td></tr><tr><td>221,8: 242,0</td><td></td><td>40</td><td>0</td></tr></table>	Метка	Доля	Кол-во	%	-1,3: 15,0		5116	51	15,0: 39,3		1530	15	39,3: 59,6		748	7	59,6: 79,8		533	5	79,8: 100,1		550	6	100,1: 120,4		420	4	120,4: 140,7		318	3	140,7: 160,9		202	2	160,9: 181,2		141	1	181,2: 201,5		110	1	201,5: 221,8		66	1	221,8: 242,0		40	0	Накладные расходы в более чем половине случаев составляют до 20 ден.ед.																																						
Метка	Доля	Кол-во	%																																																																																									
-1,3: 15,0		5116	51																																																																																									
15,0: 39,3		1530	15																																																																																									
39,3: 59,6		748	7																																																																																									
59,6: 79,8		533	5																																																																																									
79,8: 100,1		550	6																																																																																									
100,1: 120,4		420	4																																																																																									
120,4: 140,7		318	3																																																																																									
140,7: 160,9		202	2																																																																																									
160,9: 181,2		141	1																																																																																									
181,2: 201,5		110	1																																																																																									
201,5: 221,8		66	1																																																																																									
221,8: 242,0		40	0																																																																																									
returns Возвраты	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>0</td><td></td><td>8886</td><td>89</td></tr><tr><td>3</td><td>1</td><td></td><td>1114</td><td>11</td></tr><tr><td>4</td><td><null></td><td></td><td>0</td><td>0</td></tr></table>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	0		8886	89	3	1		1114	11	4	<null>		0	0	После замены и приведения к бинарному виду: 11% (1114 из 10000) были возвраты.																																																																	
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	0		8886	89																																																																																								
3	1		1114	11																																																																																								
4	<null>		0	0																																																																																								
Sales Продажи	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td><null></td><td></td><td>0</td><td>0</td></tr><tr><td>2</td><td>6:1963</td><td></td><td>6398</td><td>64</td></tr><tr><td>3</td><td>1963:3926</td><td></td><td>2609</td><td>26</td></tr><tr><td>4</td><td>3926:5889</td><td></td><td>346</td><td>3</td></tr><tr><td>5</td><td>5889:7852</td><td></td><td>58</td><td>1</td></tr><tr><td>6</td><td>7852:9815</td><td></td><td>118</td><td>1</td></tr><tr><td>7</td><td>9815:11778</td><td></td><td>71</td><td>1</td></tr><tr><td>8</td><td>11778:13741</td><td></td><td>33</td><td>0</td></tr><tr><td>9</td><td>13741:15704</td><td></td><td>29</td><td>0</td></tr><tr><td>10</td><td>15704:17667</td><td></td><td>48</td><td>0</td></tr><tr><td>11</td><td>17667:19630</td><td></td><td>20</td><td>0</td></tr><tr><td>12</td><td>19630:21593</td><td></td><td>11</td><td>0</td></tr></table>	№	Метка	Доля	Кол-во	%	1	<null>		0	0	2	6:1963		6398	64	3	1963:3926		2609	26	4	3926:5889		346	3	5	5889:7852		58	1	6	7852:9815		118	1	7	9815:11778		71	1	8	11778:13741		33	0	9	13741:15704		29	0	10	15704:17667		48	0	11	17667:19630		20	0	12	19630:21593		11	0	Расчет по формуле: (Quantity * Price). 90% продаж приходится на сумму до 1960 ден.ед. Столбец будет использован для формирования прогноза продаж на последующий квартал.																									
№	Метка	Доля	Кол-во	%																																																																																								
1	<null>		0	0																																																																																								
2	6:1963		6398	64																																																																																								
3	1963:3926		2609	26																																																																																								
4	3926:5889		346	3																																																																																								
5	5889:7852		58	1																																																																																								
6	7852:9815		118	1																																																																																								
7	9815:11778		71	1																																																																																								
8	11778:13741		33	0																																																																																								
9	13741:15704		29	0																																																																																								
10	15704:17667		48	0																																																																																								
11	17667:19630		20	0																																																																																								
12	19630:21593		11	0																																																																																								
Expenses Закупочная ст-ть за объем товара	Расчет по формуле: (Quantity * Cost)																																																																																											
net profit Чистая прибыль	Расчет по формуле: (Sales – Expenses – OverHead)																																																																																											
net margin Чистая маржа	Расчет по формуле: (net profit / Sales)																																																																																											

Так как между экономическими показателями из таблицы data есть прямая зависимость, необходимо удалить столбцы, которые могут быть рассчитаны по средствам других столбцов. А именно предлагаю удалить: Price, Expenses, net profit, net margin.

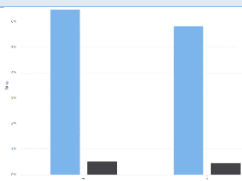
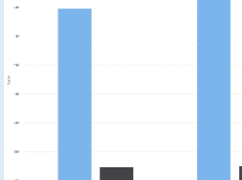
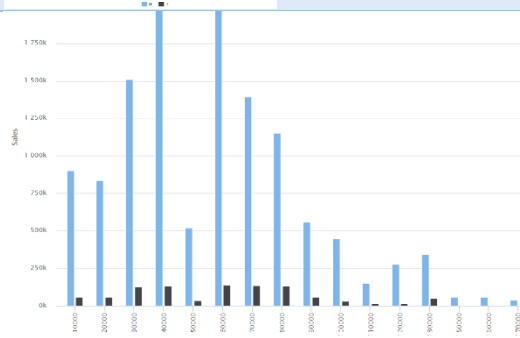
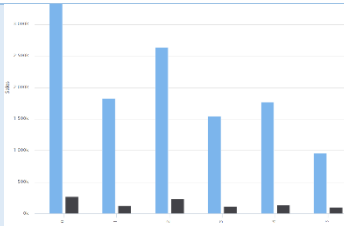
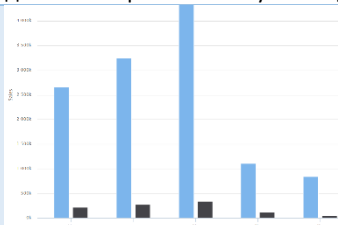
Лист 2 – products – до объединения содержал 11 столбцов и 606 строк. Это вся номенклатура товаров, которые продает данный магазин. Товары разделены на категории (4 штуки) и подкатегории (38 штук). Изучим данный лист более подробно:

Наименование	Планируемые действия																																																											
ProductKey №товара	Содержит 606 уникальных значений. Не будем учитывать.																																																											
Product Наименование товара	Данный столбец добавим к основному массиву данных (таблица data).																																																											
Product Sub Category Подкатегория товара	<div></div> <p>38 подкатегорий. Судя по статистике данного признака чаще всего возвращают велосипеды типа: Touring, Road, Mountain, а также рамы. Возвраты по остальным подкатегориям минимальны. Считаю, что признак важен для построения модели.</p>																																																											
Product Category Категория товара	<div></div> <p>4 категории: Велосипеды, Экипировка, Комплектующие, Аксессуары. Если рассматривать возвраты в разрезе категорий, то «ответственность» отдельных типов товара размывается. А так как признаки категория и подкатегория сильно коррелируют, предлагаю удалить столбец «Категория».</p>																																																											
Color Цвет	<table><thead><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr></thead><tbody><tr><td>1</td><td>Black</td><td><div></div></td><td>3062</td><td>31</td></tr><tr><td>2</td><td>Blue</td><td><div></div></td><td>725</td><td>7</td></tr><tr><td>3</td><td>Grey</td><td><div></div></td><td>40</td><td>0</td></tr><tr><td>4</td><td>Multi</td><td><div></div></td><td>527</td><td>5</td></tr><tr><td>5</td><td>NA</td><td><div></div></td><td>1636</td><td>16</td></tr><tr><td>6</td><td>Red</td><td><div></div></td><td>1545</td><td>15</td></tr><tr><td>7</td><td>Silver</td><td><div></div></td><td>1064</td><td>11</td></tr><tr><td>8</td><td>Silver/Black</td><td><div></div></td><td>137</td><td>1</td></tr><tr><td>9</td><td>White</td><td><div></div></td><td>118</td><td>1</td></tr><tr><td>10</td><td>Yellow</td><td><div></div></td><td>1146</td><td>11</td></tr></tbody></table> <div></div> <p>Признак с точки зрения возвратов важный.</p>					№	Метка	Доля	Кол-во	%	1	Black	<div></div>	3062	31	2	Blue	<div></div>	725	7	3	Grey	<div></div>	40	0	4	Multi	<div></div>	527	5	5	NA	<div></div>	1636	16	6	Red	<div></div>	1545	15	7	Silver	<div></div>	1064	11	8	Silver/Black	<div></div>	137	1	9	White	<div></div>	118	1	10	Yellow	<div></div>	1146	11
№	Метка	Доля	Кол-во	%																																																								
1	Black	<div></div>	3062	31																																																								
2	Blue	<div></div>	725	7																																																								
3	Grey	<div></div>	40	0																																																								
4	Multi	<div></div>	527	5																																																								
5	NA	<div></div>	1636	16																																																								
6	Red	<div></div>	1545	15																																																								
7	Silver	<div></div>	1064	11																																																								
8	Silver/Black	<div></div>	137	1																																																								
9	White	<div></div>	118	1																																																								
10	Yellow	<div></div>	1146	11																																																								
Size Размер	26% данных не опознаны (N/A). Также нет высокой корреляции с возвратами. Удалим столбец.																																																											
Weight Вес	128 уникальных значений. Удалим столбец.																																																											

Class Класс	 <p>Столбец важен для анализа, т.к. есть взаимосвязь между классом и возвратами.</p>
Style Стиль	 <p>Столбец важен для анализа, т.к. есть взаимосвязь между классом и возвратами.</p>
ModelName Модель	120 уникальных значений. Это вариант группировки товаров по их моделям. По сути наименование товара - это модель + цвет + размер. Можно из анализа убрать точное наименование продукта, а оставить модель и ее характеристики. Вопрос вызывает только значение NULL (нет модели), но т.к. под него попадает 409 транзакций (4%), посчитаем этот факт незначительным. Удалим Product.
StandardCost Затраты	На первый взгляд данный столбец должен коррелировать со столбцом Cost таблицы data. Но это не так – данные не совпадают. С учетом того, что данные о закупочной стоимости есть в data, данный столбец не будем использовать.

Лист 3 и 4 – customers – до объединения содержал 13 столбцов и 9586 строк. Пропусков нет. Это список обращений покупателей. Данная таблица дополнена данными из таблицы **customerProfile** - список профилей покупателей (12 столбцов и 1000 строк, пропусков нет). Необходимо объединить данные из этих таблиц. Изучим листы:

Наименование	Планируемые действия																																										
customerKey №клиента	9586 уникальных значений.																																										
ProfileKey №профиля	1000 уникальных значений.																																										
FirstName Имя	Для дальнейшего анализа столбец Имя и Фамилия мы объединим и добавим к основному массиву данных (таблица data). Что удивительно – уникальных имен 9586. А профилей только 1000. Приходим к выводу, что профили создаются для групп лиц, подходящих под определенные критерии. Поэтому номер профиля необходимо оставить для дальнейшего анализа.																																										
LastName Фамилия																																											
Address Адрес	9077 уникальных значений. Удалим столбец.																																										
City Город	2344 уникальных значения. Удалим столбец.																																										
County Округ	880 уникальных значений. Может понадобится при анализе продаж в определенном штате. Но на текущий момент столбец считаю необходимым удалить.																																										
State code Код штата	79 уникальных значений. Кодирует и соответственно дублирует информацию столбца State. Удалим столбец.																																										
State Штат	79 уникальных значений. Важно отметить, что штаты выделены во всех рассмотренных странах. Данные необходимо добавить к основному массиву.																																										
Phone Телефон	9123 уникальных значения. Не информативен для анализа. Удалим столбец.																																										
Email Эл.почта	9585 уникальных значений. Не информативен для анализа. Удалим столбец.																																										
Country Страна	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th><th></th></tr><tr><td>1</td><td>Australia</td><td><div></div></td><td>323</td><td>3</td><td rowspan="6">6 стран: Австралия, Канада, Франция, Германия, Англия, США. 90% продаж приходится на США. Столбец не информативен сам по себе.</td></tr><tr><td>2</td><td>Canada</td><td><div></div></td><td>103</td><td>1</td></tr><tr><td>3</td><td>France</td><td><div></div></td><td>194</td><td>2</td></tr><tr><td>4</td><td>Germany</td><td><div></div></td><td>162</td><td>2</td></tr><tr><td>5</td><td>United Kingdom</td><td><div></div></td><td>187</td><td>2</td></tr><tr><td>6</td><td>United States</td><td><div></div></td><td>9031</td><td>90</td></tr></table>						№	Метка	Доля	Кол-во	%		1	Australia	<div></div>	323	3	6 стран: Австралия, Канада, Франция, Германия, Англия, США. 90% продаж приходится на США. Столбец не информативен сам по себе.	2	Canada	<div></div>	103	1	3	France	<div></div>	194	2	4	Germany	<div></div>	162	2	5	United Kingdom	<div></div>	187	2	6	United States	<div></div>	9031	90
№	Метка	Доля	Кол-во	%																																							
1	Australia	<div></div>	323	3	6 стран: Австралия, Канада, Франция, Германия, Англия, США. 90% продаж приходится на США. Столбец не информативен сам по себе.																																						
2	Canada	<div></div>	103	1																																							
3	France	<div></div>	194	2																																							
4	Germany	<div></div>	162	2																																							
5	United Kingdom	<div></div>	187	2																																							
6	United States	<div></div>	9031	90																																							
ZIP Индекс	4381 уникальных значений. Некоторые значения состоят из 4 цифр – это не правильный формат. Считаю, что данный столбец необходимо исключить.																																										

Наименование		Планируемые действия																															
profileKey	№профиля	1000 уникальных значений.																															
Marital Status	Семейное положение		Замужем (женат) – 52% или Одинок(а) – 48%. Выборка в целом равномерная. Но данный признак не оказывает влияния на возвраты: и женатые и свободные делают возвраты в 8% случаев. Признак удалим.																														
Gender	Пол		Мужчина – 52% или Женщина – 48%. Выборка равномерная. Данный признак не оказывает серьезного влияния на возвраты: и женщины лишь на 0,004 п.п. чаще делают возврат. Признак удалим.																														
Income	Доход		От 10 000 до 170 000 с шагом 10 000 (16 корзин). Т.к. разбег цифр огромный предлагаю сделать замену значений Например: 10000->10, 170000->17. Это уменьшит разброс. Люди со средним доходом 40-60 тыс.ден.ед чаще совершают покупки, но и чаще делают возвраты. Люди с доходом более 150 т.д.е. возвраты не делают.																														
Children	Количество детей		От 0 до 5. Интересно отметить, что в данный магазин обращаются по большей части люди с детьми – 72% имеют хотя бы 1 ребенка. Если оценить процент возврата в зависимости от объема продаж, то мы увидим, что количество детей не влияет, возвраты по каждой группе 7-9 процентов. Удалим признак.																														
Education	Образование	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td>Bachelors</td><td><div></div></td><td>2969</td><td>30</td></tr><tr><td>2</td><td>Graduate Degree</td><td><div></div></td><td>1697</td><td>17</td></tr><tr><td>3</td><td>High School</td><td><div></div></td><td>1842</td><td>18</td></tr><tr><td>4</td><td>Partial College</td><td><div></div></td><td>2711</td><td>27</td></tr><tr><td>5</td><td>Partial High Sc...</td><td><div></div></td><td>781</td><td>8</td></tr></table>	№	Метка	Доля	Кол-во	%	1	Bachelors	<div></div>	2969	30	2	Graduate Degree	<div></div>	1697	17	3	High School	<div></div>	1842	18	4	Partial College	<div></div>	2711	27	5	Partial High Sc...	<div></div>	781	8	5 вариантов: Бакалавр, Высшее образование, Средняя школа, Неоконченный колледж, Неоконченная средняя школа. Получается, что 47% - имеют высшее образование, по ним возвраты 10%. Люди без высшего образования (53%) возвраты делают только в 6% случаев. Признак важен, но необходимо сделать его бинарным – есть или нет высшее образование.
№	Метка	Доля	Кол-во	%																													
1	Bachelors	<div></div>	2969	30																													
2	Graduate Degree	<div></div>	1697	17																													
3	High School	<div></div>	1842	18																													
4	Partial College	<div></div>	2711	27																													
5	Partial High Sc...	<div></div>	781	8																													
Occupation	Род занятий	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td>Clerical</td><td><div></div></td><td>1758</td><td>18</td></tr><tr><td>2</td><td>Management</td><td><div></div></td><td>1649</td><td>16</td></tr><tr><td>3</td><td>Manual</td><td><div></div></td><td>1227</td><td>12</td></tr><tr><td>4</td><td>Professional</td><td><div></div></td><td>2758</td><td>28</td></tr><tr><td>5</td><td>Skilled Manual</td><td><div></div></td><td>2608</td><td>26</td></tr></table>	№	Метка	Доля	Кол-во	%	1	Clerical	<div></div>	1758	18	2	Management	<div></div>	1649	16	3	Manual	<div></div>	1227	12	4	Professional	<div></div>	2758	28	5	Skilled Manual	<div></div>	2608	26	5 вариантов: Клерк, Управление, Физический труд, Специалист, Квалифицированный рабочий. Люди разных профессий примерно в одинаковых условиях по возвратам 7-9% по каждой группе.
№	Метка	Доля	Кол-во	%																													
1	Clerical	<div></div>	1758	18																													
2	Management	<div></div>	1649	16																													
3	Manual	<div></div>	1227	12																													
4	Professional	<div></div>	2758	28																													
5	Skilled Manual	<div></div>	2608	26																													
Home Owner	Недвижимость	Да – 68% или Нет – 32%. Но признак не влияет на возвраты, и та и другая группа делает возвраты в 8% случаев. Удалим.																															
Cars	Количество машин		От 0 до 4. Зависимость с точки зрения возвратов есть: 0 или 1 машина – 8% 2 машины – 7% 3 машины – 10% 4 машины – 5%																														
Commute Distance	Отдаленность	<table><tr><th>№</th><th>Метка</th><th>Доля</th><th>Кол-во</th><th>%</th></tr><tr><td>1</td><td>0-1 Miles</td><td><div></div></td><td>3615</td><td>36</td></tr><tr><td>2</td><td>10+ Miles</td><td><div></div></td><td>1095</td><td>11</td></tr><tr><td>3</td><td>1-2 Miles</td><td><div></div></td><td>1720</td><td>17</td></tr><tr><td>4</td><td>2-5 Miles</td><td><div></div></td><td>1639</td><td>16</td></tr><tr><td>5</td><td>5-10 Miles</td><td><div></div></td><td>1931</td><td>19</td></tr></table>	№	Метка	Доля	Кол-во	%	1	0-1 Miles	<div></div>	3615	36	2	10+ Miles	<div></div>	1095	11	3	1-2 Miles	<div></div>	1720	17	4	2-5 Miles	<div></div>	1639	16	5	5-10 Miles	<div></div>	1931	19	Явно выделяется группа с отдаленностью места жительства до 1 мили. Что логично – эта группа людей, может заменить передвижение на машине, ездой на велосипеде. Но сделать признак бинарным мы не можем, т.к. у каждой группы различное влияние на возвраты: 0-1 Miles – 8,5%, 1-2 Miles – 6,8%, 2-5 Miles – 9,6%, 5-10 Miles – 6%, 10+ Miles – 10%.
№	Метка	Доля	Кол-во	%																													
1	0-1 Miles	<div></div>	3615	36																													
2	10+ Miles	<div></div>	1095	11																													
3	1-2 Miles	<div></div>	1720	17																													
4	2-5 Miles	<div></div>	1639	16																													
5	5-10 Miles	<div></div>	1931	19																													

Age Возраст	№	Метка	Доля	Кол-во	%	53 уникальных значения: от 25 до 89. Товары данного магазина более всего популярны среди людей средних лет. Ввиду большого количества уникальных значений округлим возраст до ближайшего десятка.
	1	<null>		0	0	
	2	25: 35		2578	26	
	3	36: 46		3438	34	
	4	47: 56		2428	24	
	5	57: 67		1277	13	
	6	68: 78		256	3	
	7	79: 89		23	0	
	8	<null>		0	0	
Purchased Bike Велосипед		Да или Нет – признак сбалансирован. Но он не оказывает большого влияния на возвраты. У тех у кого есть велосипед возвраты в 8,3% случаев. У кого нет – 7,7%.				

Лист 5 и 6 – **manufacturers** – 2 столбца и 11 строк, пропусков нет. Это список производителей. Данная таблица дополнена данными из таблицы **manufacturer details** – подробности о производителях (3 столбца и 11 строк). Изучим данные листы:

Наименование	Планируемые действия
manfKey № производителя	
Manufacturer Наименование производителя	Данный столбец добавим к основному массиву данных (таблица data). Удалив при этом поле data.manfKey

Наименование	Планируемые действия
Manufacturer Наименование производителя	
Manufacturer Grouping Группа производителей	 <p>Новички – 64% или Действующие лица – 36%. Признак оказался очень важным: По новичкам возвраты делают только 5,8% случаев, не смотря на высокие продажи по данной группе. А вот по действующим игрокам возвраты в 12,5% случаев.</p>
Market Рынок	 <p>Стабильный – 54%, возвраты в 6,5% случаев, Растущий – 27% возвраты в 9,5% случаев, Снижающийся – 18% возвраты в 10,7% случаев. Признак важный.</p>

Лист 7 – promotions – 4 столбца и 16 строк, пропусков нет. Это список акций, которые проводит магазин. Выше мы уже приняли решение о незначительности данного признака и удалении всей информации, связанной с промо-акциями. Поэтому углубляется в анализ данной таблицы не будем.

3. Преобразование данных кейса

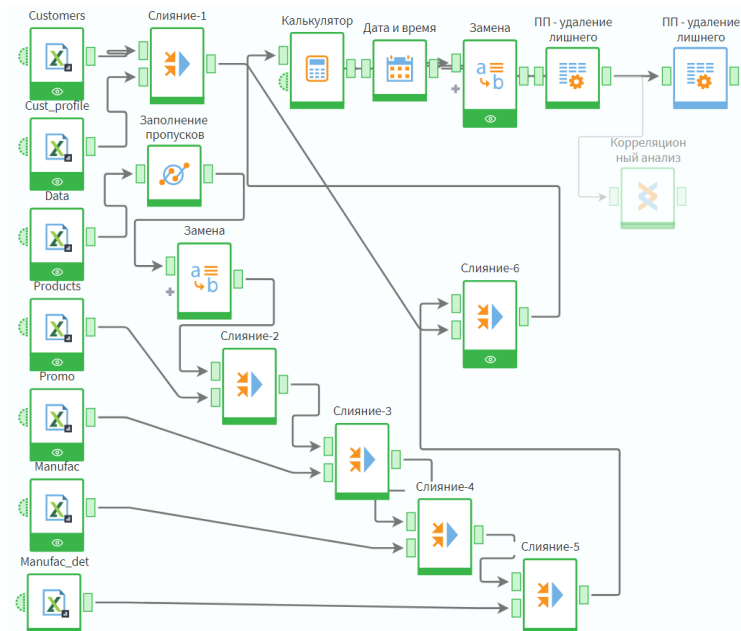
Все запланированные выше преобразования проведем с помощью инструментария платформы Loginom. Схема работ ниже:

- 1) С помощью блока «Калькулятор»:
 - a. соединим поля Имя и Фамилия (`Concat(FirstName, " ", LastName)`)
 - b. округлим до ближайшего десятка возраст (`Round(Age,-1)`)
 - c. отобразим поле Доход в тысячах ден.единиц (`Income/1000`)
- 2) С помощью блока «Дата и время» добавим данные формата «Год+Месяц».
- 3) С помощью блока «Замена» сделаем признак Образование бинарным. Если Bachelors и Graduate Degree, то Yes. Если остальное, то No.
- 4) С учетом проведенного выше анализа удалим лишние столбцы с помощью блока «Параметры полей». В этом же блоке отредактируем имена полей, в которых есть пробелы. Новые столбцы переименуем по смыслу.
- 5) Проведем корреляционный анализ с помощью одноименного блока. Высокая корреляция:
 - a. Quantity: 0.77 Sales
 - b. Cost: 0.73 OverHead, 0.70 Sales, 0.57 ModelName
 - c. OverHead: 0.52 Sales, 0.45 ModelName
 - d. Manufacturer Grouping: 0.59 Manufacturer

Для повышения качества модели машинного обучения удалим столбцы Cost, OverHead, Manufacturer Grouping.

Столбец Количество нам необходим для проведения дальнейшего анализа, а также для агрегации данных. Его удалять мы не будем.

Ниже на рисунках представлены схема консолидации и подготовки данных в Loginom, а также вывод данных сформированной таблицы (21 столбца).



1	ab	FullName	Reggie Bernet
2	31	Date	28.06.2008, 00:00
3	31	Date_YM	01.06.2008, 00:00
4	ab	Distance	10+ Miles
5	12	RoundAge	30
6	12	Income_K	110
7	ab	Manufacturer	Slicenger
8	ab	row	076CE773-4C1C-4596-8A87-0000371819E9
9	12	Quantity	1
10	90	Sales	1 315,00
11	ab	Product Sub Cate...	Road Bikes
12	ab	Color	Red
13	ab	Class	H
14	ab	Style	U
15	ab	ModelName	Road-250
16	ab	Market	Stable
17	12	ProfileKey	423
18	ab	State	New York
19	12	Cars	3
20	ab	HighEducation	No
21	12	Returns	0

4. Обогащение данных кейса. ABC, XYZ, RFM анализ клиентов

1) С помощью блока «Группировка» создаем дополнительные столбцы с данными агрегированными по полю Имя: Sales (Сумма), Quantity (Сумма, Среднее, Стандартное отклонение), Date (Максимальное), Row (Количество).

2) ABC-анализ (Принцип Парето 80/15/5):

- Сгруппируем данные по полю Sales (блок «Группировка»)
- Рассчитаем долю продаж по конкретному клиенту в общей сумме продаж $((Sales/Stat("Sales", "Sum")) * 100)$. Назовем столбец part_sales.
- Рассчитаем накопленную долю продаж по клиентам отсортированным по убыванию (т.е. начиная от лучших клиентов). Назовем столбец sum_part. Формула - $CumulativeSum("part_sales")$
- Зададим условия точно согласно принципу Парето, что если накопленная доля продаж не превышает 80%, то клиент «А»; если доля от 80-95%, то клиент «В», более 95% - клиент «С». Формула - $IF(sum_part \leq 80, "A", IF(sum_part > 95, "C", "B"))$

№	Метка	Значение
1	9.0 part_sales	0,30
2	9.0 sum_part	0,30
3	ab ABC	A
4	ab FullName	Stanley Weber
5	9.0 Sales Сумма	39 652,41
6	9.0 Quantity Сумма	9,00
7	9.0 Quantity Среднее	3,00
8	9.0 Quantity Стандартное откл.	3,46
9	31 Date Максимум	27.03.2010, 00:00
10	12 row Количество	3

№	Метка	Доля	Кол-во	%
1	A	<div></div>	2305	24
2	B	<div></div>	1922	20
3	C	<div></div>	5358	56

Наши данные практически идеально отображают принцип Парето. 80% дохода приносят 24% клиентов. 15% дохода - 20% клиентов и оставшиеся 56% клиентов – только 5% продаж.

3) XYZ-анализ (стабильность потребления: постоянно, сезонно, редко):

- Посчитаем коэффициент вариации, как $Quantity_Std/Quantity_Avg * 100$.
- Зададим условие $IF(VAR \leq 10, "X", IF(VAR > 25, "Z", "Y"))$. При коэффициенте вариации менее 10 – клиент «X», от 10 до 25 «Y», более 25 «Z».

№	Метка	Значение
1	9.0 VAR	0,00
2	ab XYZ	X
3	ab FullName	Reggie Bernet
4	9.0 Sales Сумма	1 315,00
5	9.0 Quantity Сумма	1,00
6	9.0 Quantity Среднее	1,00
7	9.0 Quantity Станда...	0,00
8	31 Date Максимум	28.06.2008, 00:00
9	12 row Количество	1

№	Метка	Доля	Кол-во	%
1	X	<div></div>	9520	99
2	Y	<div></div>	1	0
3	Z	<div></div>	64	1

В конкретно нашем случае XYZ-анализ бесполезен. Давайте рассмотрим статистику данных после группировки по клиентам: 9585 уникальных записей (до группировки было 10000). Делаем ввод, что менее 415 клиентов приходили повторно. Соответственно стандартное отклонение в таких условиях стремится к нулю в 99% случаев $((1 \text{ покупка} - 1 \text{ покупка в среднем}) / 1 = 0)$. Что приводит к нулевому коэффициенту вариации $(0/1=0)$ также в 99% случаев. Смотрим статистику коэффициента вариации:

№	Метка	Доля	Кол-во	%
1	<null>		0	0
2	0,00000: 0,00115	<div></div>	9520	99
3	0,00115: 0,00231	<div></div>	0	0
4	0,00231: 0,00346	<div></div>	0	0

Получается, что 99% клиентов – это X-клиенты. Для анализа эта информация бессмысленна. Поэтому добавлять ее в основной массив данных не будем.

4) RFM-анализ (анализ клиента по давности, приносимому им доходу и частоте обращений):

- Рассчитаем поле Recency (свежесть или давность). Показатель рассчитывается как разница между определённой датой (сегодня, конец года ли что-то другое) и максимальной датой заказа. `DaysBetween(Today(),Date)`
- Далее необходимо поделить всех клиентов на 3 группы (можно и 5, но мы разделим на 3) с помощью блока «Квантование» по следующим показателям:
 - Давность: давно был – 1, недавно – 3, посередине – 2
 - Сумма покупок: большая сумма – 3, маленькая – 1, посередине – 2
 - Частота обращения: много заказов – 3, мало заказов -1, посередине – 2
- Затем с помощью «Калькулятора» соединим эти 3 составляющие. `Concat(Recency_Label,row_Label, Sales_Label)`

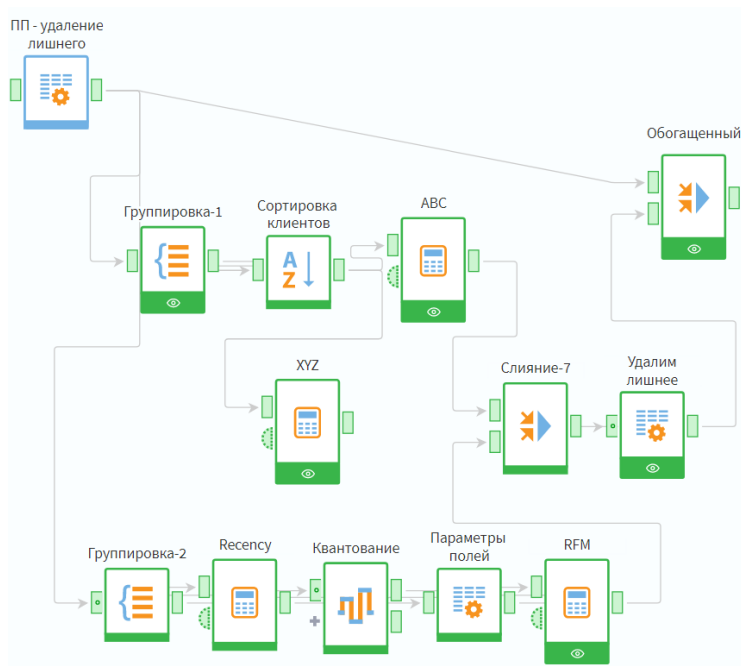
№	Метка	Доля	Кол-во	%
1	111		1408	15
2	112		69	1
3	113		7	0
4	211		5072	53
5	212		111	1
6	213		14	0
7	221		19	0
8	222		1	0
9	223		1	0
10	232		1	0
11	311		2765	29
12	312		54	1
13	313		12	0
14	321		45	0
15	322		4	0
16	323		1	0
17	332		1	0

97% операций приходится на 3 группы:

- 111 (15%) – клиент был давно, потратил мало, приходит редко
- 211 (53%) – был не так давно, потратил мало, приходит редко
- 311 (29%) – был недавно, потратил мало, приходит редко

Такая статистика говорит о том, что в данный магазин приходят в основном за разовой покупкой и чаще всего больше не возвращаются.

5) Далее необходимо обогатить наш датасет данными ABC, RFM анализа. Это можно сделать с помощью слияния. Схема работы и результат приведены ниже.

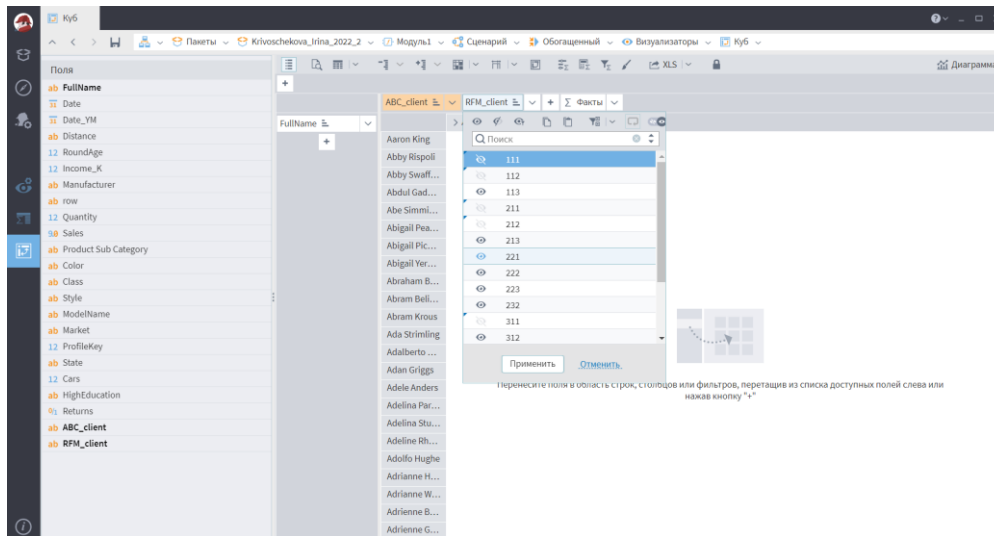


1	ab	FullName	Reggie Bernet
2	31	Date	28.06.2008, 00:00
3	31	Date_YM	01.06.2008, 00:00
4	ab	Distance	10+ Miles
5	12	RoundAge	30
6	12	Income_K	110
7	ab	Manufacturer	Slicenger
8	ab	row	076CE773-4C1C-4596-8A87-0000371819E9
9	12	Quantity	1
10	90	Sales	1 315,00
11	ab	Product Sub Cate...	Road Bikes
12	ab	Color	Red
13	ab	Class	H
14	ab	Style	U
15	ab	ModelName	Road-250
16	ab	Market	Stable
17	12	ProfileKey	423
18	ab	State	New York
19	12	Cars	3
20	ab	HighEducation	No
21	12	Returns	0
22	ab	ABC_client	B
23	ab	RFM_client	111

- Теперь мы готовы дать ответ на задачу №3 «Провести анализ клиентов, составить список приоритетных клиентов для формирования персональных предложений. Можно ли составить портрет «идеального» клиента?».

В этот список попадут все клиенты с категорией «А» из ABC-анализа – это 2305 человек, которые формируют 80% дохода магазина. Также в список попадут лучшие клиенты из RFM-анализа из следующих категорий: 113, 213, 221, 222, 223, 232, 312, 313, 321, 322, 323, 332 – итого 160 человек.

- 7) Для того чтобы выгрузить список «полезных» клиентов воспользуемся инструментом «Куб», в него подадим обогащенный датасет, настроим фильтры по полям ABC_client, RFM_client, которые мы наметили пунктом выше. Чтобы сформировать портрет идеального клиента в поле Фат выведем среднее по полям RoundAge, Income_K, Cars, Sales, Quantity.



	RoundAge	Income_K	Cars	Sales	Quantity	Итого:	RoundAge	Income_K	Cars
Tanya Ramos	30,00	10,00	0,00	608,26	1,00		30,00	10,00	
Tasha Freeland	30,00	30,00	1,00	17 834,17	4,00		30,00	30,00	
Tera Motes	50,00	80,00	2,00	31 367,00	7,00		50,00	80,00	
Terry Jai	50,00	130,00	0,00	1 176,06	1,00		50,00	130,00	
Toni Arun	60,00	30,00	0,00	636,54	1,00		60,00	30,00	
Tori Heryford	50,00	90,00	1,00	17 834,17	4,00		50,00	90,00	
Ty Agard	60,00	90,00	2,00	24 710,11	6,00		60,00	90,00	
Valerie Zhou	40,00	90,00	0,00	1 504,56	1,00		40,00	90,00	
Verna Fertik	30,00	10,00	1,00	32 086,56	8,00		30,00	10,00	
Vicky Browman	40,00	70,00	1,00	12 686,00	5,00		40,00	70,00	
Virginia Sara	40,00	40,00	2,00	1 261,98	1,00		40,00	40,00	
Wade Lesneski	50,00	10,00	1,00	23 667,81	6,00		50,00	10,00	
Warren Chandler	60,00	40,00	4,00	5 609,03	1,67		60,00	40,00	
Wilbert Doyon	50,00	70,00	1,00	30 476,16	6,00		50,00	70,00	
Williams Lighthall	40,00	70,00	0,00	16 386,81	8,00		40,00	70,00	
Willie Black	60,00	90,00	2,00	9 152,12	2,50		60,00	90,00	
Willie Xu	60,00	90,00	3,00	13 632,06	3,50		60,00	90,00	
Yong Zin	40,00	130,00	2,00	21 328,33	7,00		40,00	130,00	
Yvette Haine	40,00	70,00	2,00	15 234,52	5,00		40,00	70,00	
Zackary Takala	30,00	40,00	2,00	36 568,54	8,00		30,00	40,00	
Итого:	44,34	57,83	1,44	8 669,76	2,72		44,34	57,83	

Вывод: «идеальных» клиентов, которые одновременно относятся к группе «А» и принадлежат одному из классов RFM-анализа, перечисленных в пункте 6, ВСЕГО 143 ЧЕЛОВЕКА из 9 586. Их список будет приложен к работе, как Приложение №1.

Портрет «идеального» клиента: в среднем ему 45 лет, доход около 60 тыс.ден.ед., имеет 1 или 2 машины, средний чек 8670 ден.ед, покупает 2 или 3 товара.

5. Машинное обучение в платформе Loginom

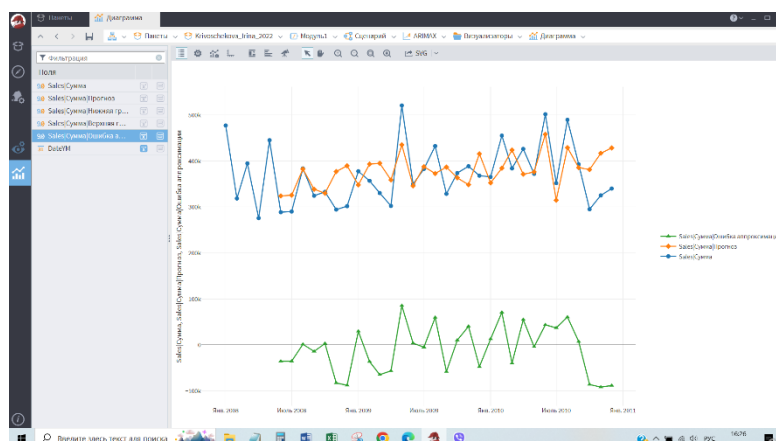
5.1. Анализ временных рядов

Ответим на первый поставленный вопрос: «Спрогнозировать продажи на следующий квартал».

С данной задачей справится анализ временного ряда, который в Loginom представлен инструментом «ARIMAX».

Для проведения анализа временных рядов необходимо наш обогащенный датасет преобразовать, сгруппировав продажи по месячно. Эти данные подадим в блок «ARIMAX». Дата помесячно - входное поле, продажи – прогнозируемое. Структуру определим автоматически, период сезонности не задаем т.к. мало данных (модель выдает ошибку), горизонт прогноза 3 месяца (на порядок ниже, чем входных данных – 36 месяцев). Обязательно рассчитаем ошибку опроксимации. Результат прогноза на ближайшие 3 месяца представлен ниже:

01.06.2010, 00:00	457 903,29
01.07.2010, 00:00	314 642,42
01.08.2010, 00:00	428 861,03
01.09.2010, 00:00	385 406,42
01.10.2010, 00:00	381 201,38
01.11.2010, 00:00	416 754,15
01.12.2010, 00:00	428 201,76
	418 087,44
	401 735,30
	449 056,31



Прогноз на последующие 3 месяца:

Период	Прогноз	Нижняя граница	Верхняя граница
Январь 2011	418 087	282 921	553 254
Февраль 2011	401 735	262 439	541 031
Март 2011	449 056	308 466	589 646

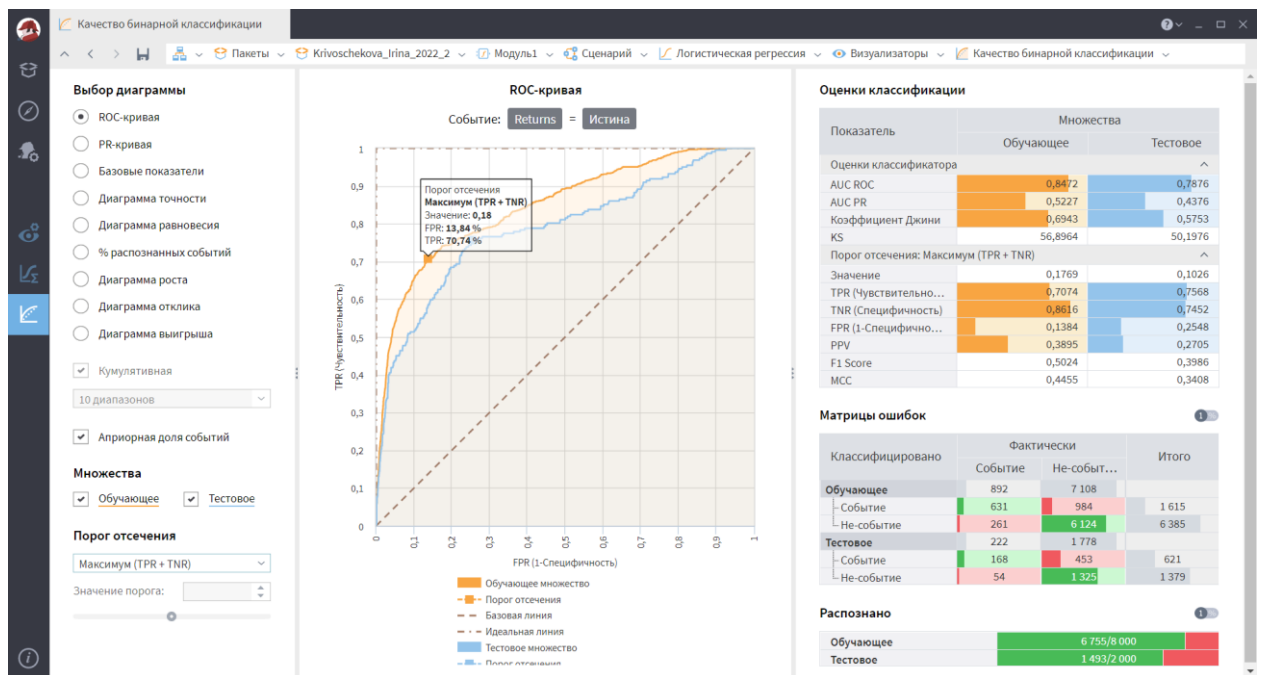
5.2. Логистическая регрессия

Ответим на второй поставленный вопрос: «Построить модель, которая будет предсказывать вероятность возврата проданного товара». Для этого выберем блок «Логистическая регрессия». Для обучения возьмем все признаки кроме Row, FullName, Date, Quantity. Выходное поле – Returns. Разобьем наш датсет на обучающее и тестовое множество, пропорция 80/20. Кросс-валидацию настраивать не будем. В остальном оставим автоматическую настройку параметров.

Настройка входных столбцов

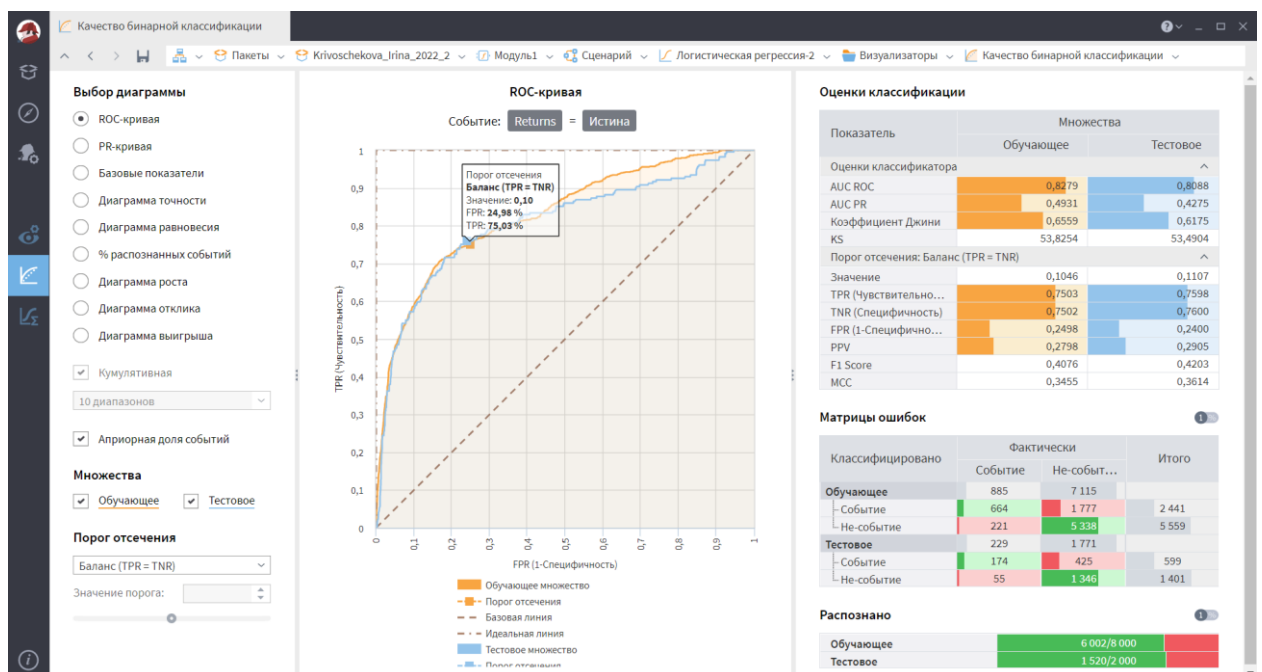
Метка	Имя	Вид данных	Назначение
ab	FullName	Дискретный	Не задано
12	Date	Непрерывный	Не задано
12	Date_YM	Непрерывный	Входное
ab	Distance	Дискретный	Входное
12	RoundAge	Непрерывный	Входное
12	Income_K	Непрерывный	Входное
ab	Manufacturer	Дискретный	Входное
ab	row	Дискретный	Не задано
12	Quantity	Непрерывный	Не задано
ab	Sales	Непрерывный	Входное
ab	Product Sub Category	Дискретный	Входное
ab	Color	Дискретный	Входное
ab	Class	Дискретный	Входное
ab	Style	Дискретный	Входное
ab	ModelName	Дискретный	Входное
ab	Market	Дискретный	Входное
12	ProfileKey	Непрерывный	Входное
ab	State	Дискретный	Входное
12	Cars	Непрерывный	Входное
ab	HighEducation	Дискретный	Входное
ab	ABC_client	Дискретный	Входное
ab	RFM_client	Дискретный	Входное
ab	Returns	Дискретный	Выходное

Чтобы посмотреть качество модели зайдём в визуализацию «Качество бинарной классификации».



Наша модель показала хорошие результаты: на тестовом множестве AUC ROC 0.79. Но F1 только 0,40. И матрица ошибок по полю «Событие» (возврат есть) показывает частые ошибки (треть случаев).

В блоке «Отчет по регрессии» изучим веса, которые модель присвоила признакам. По некоторым полям коэффициенты стремятся к нулю, удалим эти поля и запустим модель еще раз: Date, Distance, RoundAge, Income_K, Sales, Market, State, Education, ABC, ProfileKey, Cars.



Кардинально ситуация не поменялась, есть небольшое улучшение на тестовых данных: AUC ROC 0.81, F1 0,42. Но упрощение модели – это положительная тенденция. Модель построена по следующим входным параметрам: Manufacturer, Product Sub Category, Color, Class, Style, ModelName, RFM. Выход – Returns.

Далее будет рассмотрено автоматическое машинное обучение в Colab – сравним результаты.

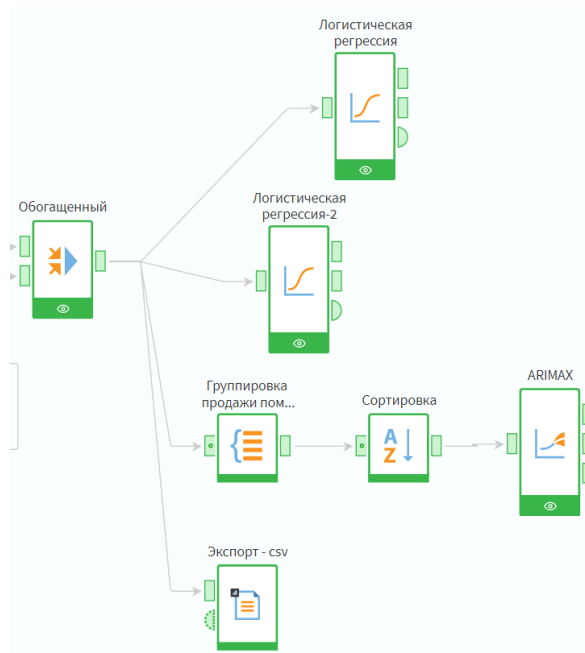


Схема по блоку «Машинное обучение в платформе Loginom»

Для дальнейшего анализа экспортируем обогащенный датасет, который был нами создан с Loginome. Блок «Экспорт – Текстовый файл». Получившийся файлы сохраним на google.диске для последующей работе в Google Colab.

6. EDA и ML в Google Colab

Дальнейшую работу продолжим в Google Colab. Проведем исследование данных на предмет выявления новых инсайдов, а также сформулируем ответы на задачи, поставленные в 1 и 2 пункте, средствами Google Colab. Обязательно сравним полученные результаты с результатами, полученными в Loginom. Ссылка на блокнот ниже:

https://colab.research.google.com/drive/1r8jaw7D8hyFyGen5T_kWGOzgaEH2la6P?usp=sharing

6.1. EDA с помощью Pandas profiling

Анализ данных, проведем с помощью библиотеки автоматической визуализации Pandas profiling. Наш датасет, выгруженный из Loginom в пункте 5.2 содержит 23 столбца и 10000 строк.

Наименование	Описание																																	
FullName Полное имя	96% - уникальные значения. Еще раз убеждаемся, что клиенты приходят в магазин в основном разово. Для построения модели лишняя информация. Удалим.																																	
Date Дата dd/mm/yyyy	Рассмотрен выше. Новых инсайдов нет.																																	
Date_YM Дата mm/yyyy	36 значений – по 12 мес. в течении 3-х лет																																	
Distance Отдаленность	Pandas profiling указывает на высокую корреляцию с такими полями как Cars, HighEducation, RoundAge, Income_K																																	
RoundAge Примерный возраст	Рассмотрен выше. Новых инсайдов нет.																																	
Income_K Доход в тыс.ден.ед.	Pandas profiling указывает на высокую корреляцию с такими полями как Cars, HighEducation, Distance																																	
Manufacturer Производители	Инсайд: По каждому производителю примерно одинаковое количество число транзакций.																																	
Row строка	Удалим.																																	
Quantity Количество	Рассмотрен выше. Новых инсайдов нет.																																	
Sales Продажи	Согласно приведенной статистике большинство значений лежит между 88 и 1514. Медиана = 237.																																	
Product_Sub_Category Подкатегория	Pandas profiling указывает на высокую корреляцию с такими полями как ABC, Style, Color.																																	
Color Цвет	<div><table><tr><th>Value</th><th>Count</th><th>Frequency (%)</th></tr><tr><td>Black</td><td>3062</td><td>30.6%</td></tr><tr><td>Red</td><td>1545</td><td>15.4%</td></tr><tr><td>Yellow</td><td>1146</td><td>11.5%</td></tr><tr><td>Silver</td><td>1064</td><td>10.6%</td></tr><tr><td>Blue</td><td>725</td><td>7.2%</td></tr><tr><td>Multi</td><td>527</td><td>5.3%</td></tr><tr><td>Silver/Black</td><td>137</td><td>1.4%</td></tr><tr><td>White</td><td>118</td><td>1.2%</td></tr><tr><td>Grey</td><td>40</td><td>0.4%</td></tr><tr><td>(Missing)</td><td>1636</td><td>16.4%</td></tr></table></div> <div>То, что в Loginom опознавалось как N/A и использовалось в анализе, здесь считается пропущенными значениями. Поэтому далее необходимо будет эти пропуски заполнить (1636 значений). Также высокая корреляция со столбцом Подкатегории.</div>	Value	Count	Frequency (%)	Black	3062	30.6%	Red	1545	15.4%	Yellow	1146	11.5%	Silver	1064	10.6%	Blue	725	7.2%	Multi	527	5.3%	Silver/Black	137	1.4%	White	118	1.2%	Grey	40	0.4%	(Missing)	1636	16.4%
Value	Count	Frequency (%)																																
Black	3062	30.6%																																
Red	1545	15.4%																																
Yellow	1146	11.5%																																
Silver	1064	10.6%																																
Blue	725	7.2%																																
Multi	527	5.3%																																
Silver/Black	137	1.4%																																
White	118	1.2%																																
Grey	40	0.4%																																
(Missing)	1636	16.4%																																
Class Класс	<div><table><tr><th>Value</th><th>Count</th><th>Frequency (%)</th></tr><tr><td>H</td><td>2813</td><td>28.1%</td></tr><tr><td>L</td><td>2672</td><td>26.7%</td></tr><tr><td>M</td><td>1757</td><td>17.6%</td></tr><tr><td>(Missing)</td><td>2758</td><td>27.6%</td></tr></table></div> <div>Также данные по грифом N/A выделились в пропущенные значения (2758 значений). Коррелирует с полем Style.</div>	Value	Count	Frequency (%)	H	2813	28.1%	L	2672	26.7%	M	1757	17.6%	(Missing)	2758	27.6%																		
Value	Count	Frequency (%)																																
H	2813	28.1%																																
L	2672	26.7%																																
M	1757	17.6%																																
(Missing)	2758	27.6%																																
Style Стиль	<div><table><tr><th>Value</th><th>Count</th><th>Frequency (%)</th></tr><tr><td>U</td><td>6294</td><td>62.9%</td></tr><tr><td>W</td><td>960</td><td>9.6%</td></tr><tr><td>M</td><td>206</td><td>2.1%</td></tr><tr><td>(Missing)</td><td>2540</td><td>25.4%</td></tr></table></div> <div>Аналогичная проблема – N/A прочитались пропусками (2540 значений). Корреляция с Class и Подкатегории.</div>	Value	Count	Frequency (%)	U	6294	62.9%	W	960	9.6%	M	206	2.1%	(Missing)	2540	25.4%																		
Value	Count	Frequency (%)																																
U	6294	62.9%																																
W	960	9.6%																																
M	206	2.1%																																
(Missing)	2540	25.4%																																
ModelName Модель	Пропущенными опознались 409 значений заменим их на «unknown».																																	
Market Рынок	Высокая корреляция с производителем.																																	
ProfileKey №профиля	Рассмотрен в первом модуле. Отмечен Pandas profiling как признак с высокой корреляцией (коррелирует с уровнем дохода).																																	
State Штат	Рассмотрен выше. Новых инсайдов нет.																																	
Cars Машины	Высоко коррелирует с признаками Образование, Отдаленность, Уровень дохода.																																	
HighEducation Высшее образование	Высокая корреляция с признаком Машины.																																	
Returns Возвраты	Рассмотрен выше. Новых инсайдов нет.																																	
ABC_client	Признак описан выше. Важно отметить корреляцию с признаком Подкатегория.																																	
RFM_client	Признак описан выше. Очень высокая корреляция с признаком Дата.																																	

После анализа проведенного с помощью Pandas profiling выполним следующие действия:

- 1) Удалим столбцы 'FullName', 'Quantity', 'row', 'Date'. По аналогии с моделью Логистической регрессии №1 в Loginom.

```
bike = bike.drop(['FullName', 'Quantity', 'row', 'Date'], axis=1)
```

- 2) Заполним пропуски в столбцах Color, Class, Style, ModelName: N/A поменяем на 'unknown' (неизвестно).

```
bike = bike.fillna({'Color': 'unknown', 'Class': 'unknown', 'Style': 'unknown', 'ModelName': 'unknown'})
```

Теперь наши данные готовы для построения модели, которая будет предсказывать вероятность возврата товара.

6.2. Построение моделей машинного обучения с помощью PyCaret

Запустим библиотеку машинного обучения с открытым исходным кодом, которая автоматизирует весь процесс обучения модели машинного обучения – PyCaret. Была выбрана именно эта библиотека, потому что она позволяет подавать на вход модели категориальные данные, без их предварительной перекодировки в числовые данные.

Установка и запуск Pycaret:

```
!pip install PyYAML==5.4.1
!pip install dataprep
!pip install pycaret
from pycaret.utils import enable_colab
enable_colab()
from pycaret.classification import *
```

Далее необходимо разбить данные на 2 части: данные для моделирования (обучение и тестирование) и данные для предсказаний в доли 95/5:

```
bike_95pr = bike.sample(frac=0.95, random_state=1234)
bike_5pr = bike.drop(bike_95pr.index)
```

Далее запустим автоматическую подготовку данных для моделирования – укажем, какой датасет использовать, выходное поле, наименование:

```
reg_experiment = setup(bike_95pr, target = 'Returns', session_id=1234,
                      log_experiment=True, experiment_name='Bike_shop')
```

На этом этапе библиотека формирует обучающее (6649 строк) и тестовое множество (2851 строка), задает основные параметры для ML. Далее командой `compare_models()` происходит обучение основным, самым популярным моделям, которые ранжируются по качеству. Результат по нашим данным представлен ниже:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9004	0.8125	0.3149	0.6303	0.4192	0.3713	0.3986	3.162
lightgbm	Light Gradient Boosting Machine	0.8986	0.7899	0.3018	0.6159	0.4044	0.3561	0.3836	0.370
lr	Logistic Regression	0.8932	0.8027	0.2055	0.5916	0.3048	0.2615	0.3050	5.190
et	Extra Trees Classifier	0.8926	0.7725	0.1831	0.6057	0.2791	0.2387	0.2902	1.590
rf	Random Forest Classifier	0.8923	0.7771	0.1173	0.6664	0.1989	0.1704	0.2477	1.482
ada	Ada Boost Classifier	0.8883	0.8013	0.1278	0.5442	0.2068	0.1709	0.2243	0.806
lda	Linear Discriminant Analysis	0.8872	0.7994	0.3123	0.5089	0.3868	0.3287	0.3407	0.510
ridge	Ridge Classifier	0.8867	0.0000	0.0250	0.6093	0.0480	0.0391	0.1055	0.078
dummy	Dummy Classifier	0.8858	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.042
svm	SVM - Linear Kernel	0.8815	0.0000	0.0145	0.0431	0.0217	0.0108	0.0127	0.156
knn	K Neighbors Classifier	0.8732	0.5328	0.0197	0.1286	0.0341	0.0046	0.0066	0.860
dt	Decision Tree Classifier	0.8445	0.6196	0.3281	0.3222	0.3245	0.2368	0.2371	0.234
qda	Quadratic Discriminant Analysis	0.7504	0.5066	0.1908	0.2243	0.0457	0.0043	0.0242	0.356
nb	Naive Bayes	0.7252	0.7336	0.6403	0.2391	0.3478	0.2177	0.2602	0.080

Для создания модели, конечно выберем лучший из предложенных вариантов: Gradient Boosting Classifier. Чтобы создать модель необходимо:

- 1) Создать модель: `create_model('gbc')`
- 2) Улучшить модель: `tune_model(gbc)`
- 3) Завершить модель: `finalize_model(gbc)`
- 4) Предсказать на тестовых данных: `predict_model()`
- 5) Сохранить модель: `save_model()`
- 6) Загрузить модель для использования на новых данных: `load_model()`
- 7) Предсказать на новых данных: `predict_model()`
- 8) Оценить результаты!

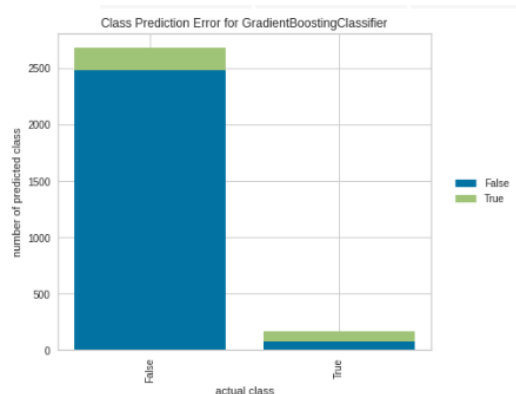
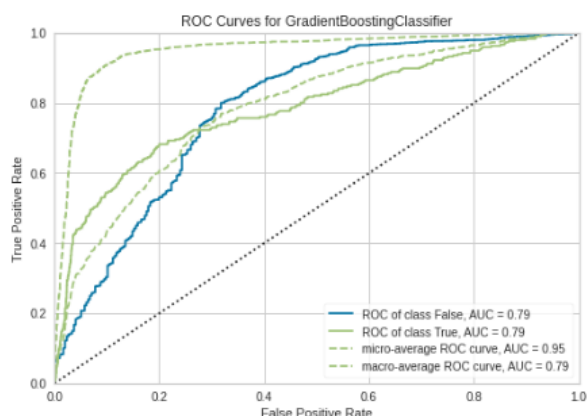
Оценим наши результаты:

Отметим, что модель Логистической регрессии от Русарет оказалась на 3 месте с результатом по AUC = 0.8027 (F1 = 0.3048), напомним, что модель от Loginom с максимальной подачей параметров показывала результат AUC = 0.79 (F1 = 0.40). Получается, что при сопоставимых результатах по AUC, модель Логистической регрессии от Loginom выигрывает по F1.

Модель Gradient Boosting Classifier, которая оказалась лучшей по версии Русарет показала следующие результаты при прогоне на скрытых данных:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	
0	Gradient Boosting Classifier	0.894	0.8129	0.3559	0.5833	0.4421	0.3873	0.4018	

	Date_YM	Distance	RoundAge	Income_K	Manufacturer	Sales	Product_Sub_Category	Color	Class	Style	...	Market	ProfileKey	State	Cars	HighEducation	Returns	ABC_client	RFM_client	Label	Score
3	2009-01-10	5-10 Miles	30	40	Ribuck	327.2400	Headsets	unknown	M	unknown	...	Stable	503	New York	1	No	False	B	211	False	0.9601
10	2010-01-11	0-1 Miles	30	50	Woolson	46.0616	Tires and Tubes	unknown	M	unknown	...	Stable	888	Louisiana	0	Yes	False	C	311	False	0.9389
15	2010-01-11	2-5 Miles	30	30	Old Balance	211.1910	Road Frames	Black	H	U	...	Shrinking	89	New South Wales	1	No	False	B	311	True	0.5151
49	2008-01-09	2-5 Miles	60	80	Nuke	1330.0000	Road Bikes	Black	L	U	...	Growing	119	Maryland	2	Yes	False	B	111	False	0.8420
54	2010-01-08	5-10 Miles	60	30	Acme	1786.0200	Mountain Bikes	Black	M	U	...	Stable	191	Utah	2	No	True	A	311	True	0.5850



Вывод: с небольшим отрывом победила модель от Русарет Gradient Boosting Classifier (метрики качества AUC = 0.8129 (F1 = 0.4421)), при этом модель Логистической регрессии от Loginom отработала лучше аналога в Русарет.

6.3. Анализ временных рядов Google Colab с помощью Prophet

Входными данными для Prophet всегда является фрейм данных с двумя столбцами: ds и y. Колонка ds (тип дата) должна быть в идеале формата ГГГГ-ММ-ДД для даты или YYYY-MM-DD HH: MM: SS для временной метки. Столбец «y» должен быть числовым, и представлять собой меру, которую мы прогнозируем.

Начнем с преобразования данных под требуемый формат. Для этого из нашего датасета выберем только Дату и Returns. Затем отсортируем их по возрастанию. И переименуем столбцы: Дата -> ds; Returns -> y. Код приведен ниже.

```
df1 = df[['Date', 'Sales']]
df1 = df1.sort_values(by=['Date'], ascending=[True])
df1.rename(columns = {'Date' : 'ds', 'Sales' : 'y'}, inplace = True)
```

Вообще Prophet можно настраивать, меняя параметры, но и с автоматическими настройками он работает хорошо, поэтому запустим базовый вариант:

```
m = Prophet()
m.fit(df1)
```

Далее необходимо указать горизонт планирования в днях. Т.к. мы планируем продажи на 3 месяца вперед, установим 90 дней:

```
future = m.make_future_dataframe(periods=90)
```

Предсказываем:

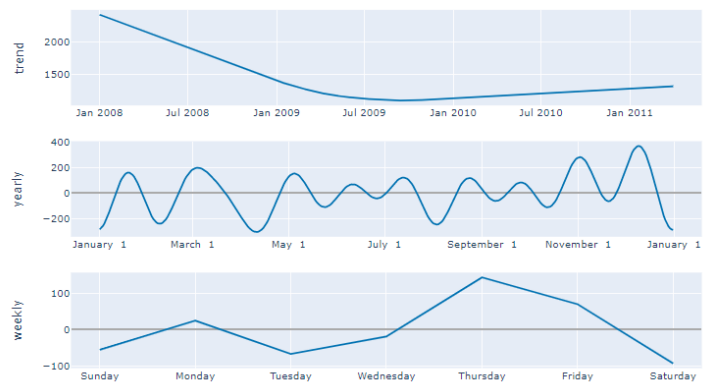
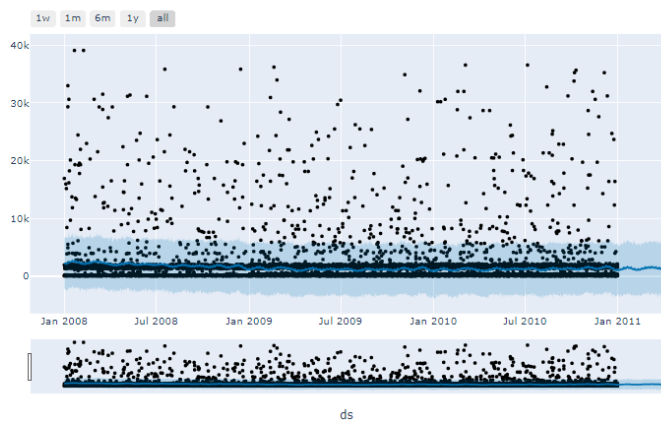
```
forecast = m.predict(future)
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail(90)
```

	ds	yhat	yhat_lower	yhat_upper
1096	2011-01-01	891.565807	-3763.332197	5172.859330
1097	2011-01-02	936.133776	-3974.881855	5350.465775
1098	2011-01-03	1029.478093	-3572.547708	5130.500785
1099	2011-01-04	956.433437	-3848.165395	5482.446801
1100	2011-01-05	1027.761967	-3470.607026	5305.169652
...
1181	2011-03-27	1167.868053	-3273.938286	5608.907433
1182	2011-03-28	1229.248222	-3250.299685	5890.315139
1183	2011-03-29	1118.307115	-3357.044787	5396.254291
1184	2011-03-30	1146.525703	-3681.419864	5628.206684
1185	2011-03-31	1290.068990	-3283.033616	5852.405806

Сохраняем результат:

```
forecast.to_csv('/content/drive/MyDrive/bikeshop3m.csv', index=False, sep=',')
```

Оценим качество работы Prophet с помощью графиков:



Судя по графикам модель, предложенная Prophet учитывает только дешевый ассортимент. Не учитывая более дорогие экземпляры. Сравним данные с результатами Loginom, а также со средними данными по этим месяцам за последние 2 года (т.к. продажи в 2009 и 2010 сопоставимы, а в 2008 г. продажи были намного меньше):

Период	Январь	Февраль	Март
Среднее за последние 2 года	371 348	405 935	356 986
Loginom	418 087	401 735	449 056
Prophet	39 048	33 562	42 588

Явно видно, что анализ временных рядов от Prophet справился намного хуже, чем Loginom. В работе будем использовать результаты Loginom.

7. Анализ ассортимента в Google BigQuery

В этом разделе мы найдем ответ на 4 поставленную задачу: «Провести анализ ассортимента: по каким товарам необходимо сформировать запас ввиду их высокого спроса, а какие товары необходимо перевести на систему предзаказа, ввиду их низкой популярности».

Для решения поставленной задачи проведем XYZ и ABC анализ ассортимента. XYZ-анализ покажет спрос на те или иные товары, стабильность продаж в определенный период, понимание, с каким постоянством продаются товары. ABC-анализ распределяет товары по группам в зависимости от прибыли, которую товары приносят магазину.

- 1) При проведении XYZ-анализа товаров рассчитывается коэффициент вариации — величина отклонения от среднего значения продаж. Весь ассортимент делится на группы:
 - в группу X относятся товары с коэффициентом вариации до 10% (спрос стабилен, товары должны быть в наличии),
 - в группу Y — товары с коэффициентом вариации 10-25% (спрос ниже, чем на товары X, их не держат на складе в большом количестве),
 - в группу Z — товары коэффициентом вариации выше 25% (спрос случайный, желательна продажа по предзаказу).
- 2) ABC-анализ распределяет товары по таким группам:
 - А — товары, которые приносят 80% дохода;
 - В — товары, которые приносят от 15% дохода;
 - С — товары, которые приносят 5% дохода.
- 3) Объединение ABC- и XYZ-анализа. Для получения более полной картины полезно совместить данные ABC и XYZ-анализ продаж, которые дополняют друг друга. Так мы объединим преимущества методов анализа и сможем подготовить советы по эффективному управлению ассортиментом.

AX	AY	AZ
Высокая прибыльность, стабильный спрос Товары должны быть на складе всегда	Высокая прибыльность, нестабильный спрос Должен быть резерв на случай спроса	Высокая прибыльность, случайный спрос Не держать на складе, но иметь возможность быстро получить от поставщика
BX	BY	BZ
Средняя прибыльность, стабильный спрос Товары должны быть на складе всегда	Средняя прибыльность, нестабильный спрос Должен быть резерв на случай спроса	Средняя прибыльность, случайный спрос Не держать на складе, но иметь возможность быстро получить от поставщика
CX	CY	CZ
Низкая прибыльность, стабильный спрос Иметь запас, исходя из обычного объема продаж	Низкая прибыльность, нестабильный спрос Создавать запас, если остался бюджет	Низкая прибыльность, случайный спрос Продавать только под заказ

Для работы нам потребуется файл с данными, которые будут содержать следующие сведения: Наименование товара (Product), Продажи (Sales). Легче всего выгрузить таблицу такого вида из Loginom в формате csv.

- 1) Создать новый датасет, в него загрузить вновь созданную таблицу «case_5_log_products», которая содержит данные о наименовании товара и сумме продаж. Таблица загрузилась на автоматическом режиме, без правок, настроила только параметр первой строки, как заголовок.
- 2) Для проведения XYZ-анализа необходимо рассчитать коэффициент вариации, который равен отношению стандартного отклонения выручки по каждому виду товара к средней выручке по этому товару. Стандартное отклонение можно получить как корень из среднеквадратичного отклонения.
Формула следующая: $\text{SQRT}(\text{STDDEV}(\text{Sales}))/\text{AVG}(\text{Sales})*100$
Также можно использовать округление: $\text{Round}(\text{SQRT}(\text{STDDEV}(\text{Sales}))*100/\text{AVG}(\text{Sales}),2)$

- 3) Создадим запрос для расчета коэффициенты вариации по каждому товару:

```
SELECT DISTINCT Product, Round(SQRT(STDDEV(Sales))*100/AVG(Sales),2) AS VAR
FROM `ira-300922.Krivoschekova_exam.products`
GROUP BY Product
```

Группировка необходима, т.к. в формуле присутствуют агрегатные функции.

- 4) Далее необходимо присвоить категории X или Y или Z в зависимости от его коэффициента вариации. Это можно сделать с помощью команды CASE, которая задает условия:

```
CASE
  WHEN a.VAR < 10 THEN 'X'
  WHEN a.VAR BETWEEN 10 AND 25 THEN 'Y'
  ELSE 'Z'
END AS XYZ
```

- 5) Теперь объединим все в единый SQL-запрос и запустим его, результаты представлены ниже:

```
1 SELECT Product, VAR,
2 CASE
3   WHEN a.VAR < 10 THEN 'X'
4   WHEN a.VAR BETWEEN 10 AND 25 THEN 'Y'
5   ELSE 'Z'
6 END AS XYZ
7 FROM (
8   SELECT DISTINCT Product, Round(SQRT(STDDEV(Sales))*100/AVG(Sales),2) AS VAR
9   FROM `ira-300922.Krivoschekova_exam.products`
10  GROUP BY Product) AS a
11 ORDER BY XYZ
```

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS	
low	Product	VAR		XYZ	
1	Stem	0.0		X	
2	Chain	0.72		X	
3	Spokes	5.32		X	
4	Decal 1	4.86		X	
5	Decal 2	6.33		X	
6	HL Fork	5.37		X	
7	LL Fork	5.02		X	

- 6) Сохраним результаты запроса как BigQuery table, чтобы в дальнейшем мы могли к ним обращаться.
- 7) Для проведения ABC-анализа необходимо рассчитать продажи по каждому виду товара, а также ту долю, которую занимают доходы от каждого товара в общем объеме продаж. Для этого создадим следующий запрос:

```
SELECT Product, sum(Sales) as PSales
FROM `ira-300922.Krivoschekova_exam.products`
GROUP BY Product
```

Результатом которого будут объемы продаж по каждому наименованию товара.

- 8) Далее создадим запрос, результатом которого будут 3 столбца, помимо наименования товара:

- Уже рассчитанные продажи по каждому товару: PSales
- Сумма накопленным итогом, начиная от самого крупного клиента (команда DESC):
`sum(PSales) OVER(ORDER BY PSales DESC)`
- Сумма всех продаж, которую можно посчитать с помощью оконной функции OVER:
`sum(PSales) OVER()`. Тогда напротив каждого товара будет одинаковая сумма общих продаж – она пригодится для дальнейших расчетов.

```
1 SELECT distinct Product, PSales,
2 sum(PSales) OVER(ORDER BY m.PSales DESC) AS Agr_sum_prod,
3 sum(PSales) OVER() AS all_Sales
4 FROM
5 (SELECT Product, sum(Sales) as PSales
6  FROM `ira-300922.Krivoschekova_exam.products`
7  GROUP BY Product
8  ORDER BY PSales) AS m
9 ORDER BY Agr_sum_prod
```

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS	
w	Product	PSales		Agr_sum_pr...	all_Sales
1	Touring-3000 Yellow, 58	598739.418...		598739.418...	13333694.9...
2	Touring-1000 Blue, 60	566697.691...		1165437.11...	13333694.9...
3	Touring-2000 Blue, 60	545194.8112		1710631.92...	13333694.9...

- 9) Создадим запрос, который будет ссылаться на ранее созданные, и будет определять категорию товара А, В или С в зависимости от места, который товар занимает в рейтинге по формированию доходов магазина. Для доход по ранжированным товарам (по величине дохода) накопленным итогом поделим на общий доход. Те товары, которые формирует 80% дохода отнесем к классу А, от 80-95% - класс В, от 95-100 – класс С. Используем уже знакомую функцию CASE. Сохраним результаты запроса как BigQuery table, чтобы в дальнейшем мы могли к ним обращаться. Запрос и его результаты представлены ниже.

```
1 SELECT Product,
2 CASE
3   WHEN Agr_sum_prod*100/all_Sales < 80 THEN 'A'
4   WHEN Agr_sum_prod*100/all_Sales BETWEEN 80 AND 95 THEN 'B'
5   ELSE 'C'
6 END AS ABC
7 FROM (SELECT distinct Product, PSales,
8       sum(PSales) OVER(ORDER BY m.PSales DESC) AS Agr_sum_prod,
9       sum(PSales) OVER() AS all_Sales
10 FROM
11 (SELECT Product, sum(Sales) as PSales
12  FROM `ira-300922.Krivoschekova_exam.products`
13  GROUP BY Product
14  ORDER BY PSales) AS m
15 ) as k
16 ORDER BY ABC
```

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS
ow	Product	ABC		
1	Road-150 Red, 48	A		
2	Mountain-200 Black, 38	A		
3	Road-650 Red, 60	A		
4	Road-650 Red, 62	A		
5	Road-650 Black, 48	A		
6	Road-550-W Yellow, 44	A		
7	Mountain-500 Silver, 44	A		
8	Tourinn-1000 Blue 46	A		

- 10) Далее необходимо объединить результаты ABC и XYZ-анализа. Это можно сделать с помощью команды CONCAT. SQL-запрос на объединение данных анализа и его результаты ниже:

```
1 SELECT DISTINCT prod.Product, CONCAT(abc.ABC, xyz.XYZ) AS ABC_XYZ
2 FROM `ira-300922.Krivoschekova_exam.products` AS prod
3 RIGHT OUTER JOIN `ira-300922.Krivoschekova_exam.abc` AS abc ON abc.Product=prod.Product
4 LEFT OUTER JOIN `ira-300922.Krivoschekova_exam.xyz` AS xyz ON prod.Product=xyz.Product
5
6
```

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS
Row	Product	ABC_XYZ		
1	Road-150 Red, 48	AX		
2	Mountain-200 Black, 38	AX		
3	Road-650 Red, 60	AX		
4	Road-650 Red, 62	AX		
5	Road-650 Black, 48	AX		
6	Road-550-W Yellow, 44	AX		

- 11) Теперь по каждому наименованию дадим рекомендацию. Для этого опять воспользуемся командой CASE:

```
1 SELECT Product, ABC_XYZ,
2 CASE
3   WHEN ABC_XYZ = 'AX' THEN 'Товары должны быть на складе всегда'
4   WHEN ABC_XYZ = 'BX' THEN 'Товары должны быть на складе всегда'
5   WHEN ABC_XYZ = 'CX' THEN 'Иметь запас исходя из обычного объема продаж'
6   WHEN ABC_XYZ = 'AY' THEN 'Должен быть резерв на случай спроса'
7   WHEN ABC_XYZ = 'BY' THEN 'Должен быть резерв на случай спроса'
8   WHEN ABC_XYZ = 'CY' THEN 'Создать запас, если остался бюджет'
9   WHEN ABC_XYZ = 'AZ' THEN 'Не держать на складе, но иметь возможность быстро получить от поставщика'
10  WHEN ABC_XYZ = 'BZ' THEN 'Не держать на складе, но иметь возможность быстро получить от поставщика'
11  WHEN ABC_XYZ = 'CZ' THEN 'Продавать только под заказ'
12 END AS Comments
13 FROM
14 (SELECT DISTINCT prod.Product, CONCAT(abc.ABC, xyz.XYZ) AS ABC_XYZ
15  FROM `ira-300922.Krivoschekova_exam.products` AS prod
16  RIGHT OUTER JOIN `ira-300922.Krivoschekova_exam.abc` AS abc ON abc.Product=prod.Product
17  LEFT OUTER JOIN `ira-300922.Krivoschekova_exam.xyz` AS xyz ON prod.Product=xyz.Product) as k
18 ORDER BY ABC_XYZ
```

Query results

JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS
ow	Product	ABC_XYZ	Comments	
1	Road-150 Red, 48	AX	Товары должны быть на скл...	
2	Mountain-200 Black, 38	AX	Товары должны быть на скл...	
3	Road-650 Red, 60	AX	Товары должны быть на скл...	
4	Road-650 Red, 62	AX	Товары должны быть на скл...	

- 12) Результаты анализа сохраним как Google Sheet, далее скачаем в виде таблицы Excel.

Анализ ассортимента с рекомендациями по складским запасам будет приложен к работе, как Приложение №2. В Excel построим сводную таблицу для анализа результатов:

ABCXYZ	Количество	Описание
AX	62	Высокая прибыльность, стабильный спрос
BX	60	Средняя прибыльность, стабильный спрос
CX	251	Низкая прибыльность, стабильный спрос
CY	35	Низкая прибыльность, нестабильный спрос
CZ	62	Низкая прибыльность, случайный спрос
Общий итог		470

7. Дашборд «Продажи по странам»

В данном пункте подготовим ответ на следующий вопрос: «Создать дашборд, отражающий продажи по странам и по периодам. Есть ли регионы, которые целесообразно покинуть?». Для этого воспользуемся двумя инструментами: Power BI, DataLens и Looker Studio (ранее Google Data Studio).

Для создания дашбордов будем использовать консолидированные данные, созданные в Loginom.

Чтобы ответить на поставленный вопрос в Power BI нам будут необходимы следующие элементы:

- 1) Матрица: демонстрация продаж и средней величины маржи по периодам;
- 2) Карточка – для демонстрации итогов по продажам, количеству и средней марже;
- 3) Карта – визуально продемонстрирует объем продаж в каждой стране и по регионам;
- 4) Текстовый блок: для подведения итогов анализа.

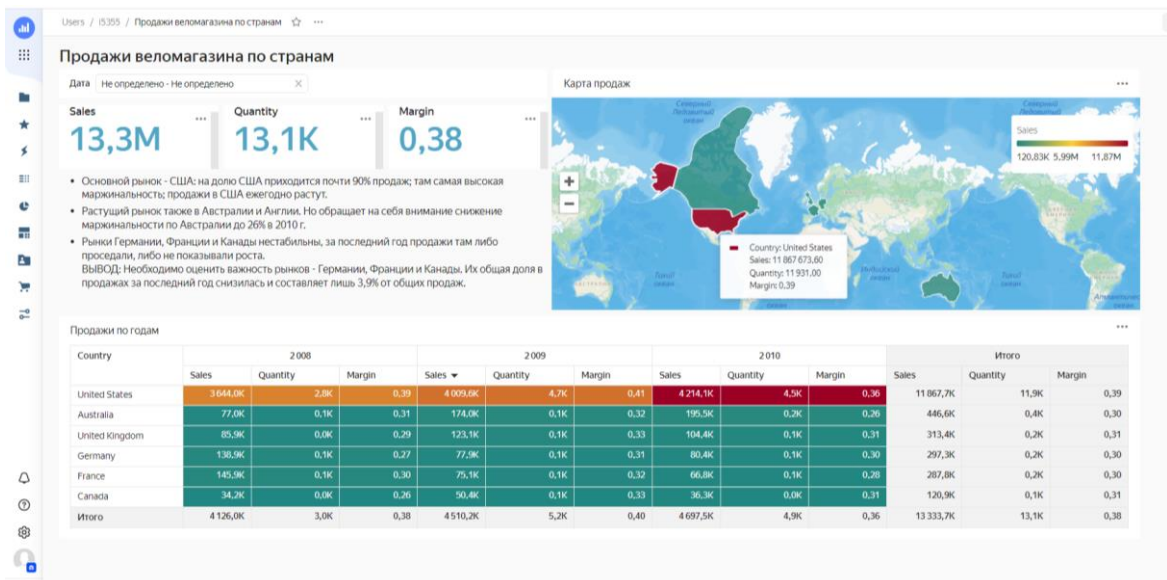
Дашборд экспортирован и будет приложен к работе как «Приложение 3. Визуализация в Power BI».



Чтобы ответить на поставленный вопрос в DataLens нам будут необходимы следующие элементы:

- 1) Селектор – позволяет выбрать необходимый интервал дат;
- 2) Индикаторы – для демонстрации итогов по продажам, количеству, средней марже;
- 3) Карта (раздел Полигоны) – визуально демонстрирует объем продаж в каждой стране, при наведении курсора мышки на страну отражаются тултипы: название страны и данные о продажах. Специально для данного чарта датасет был дополнен координатами стран;
- 4) Сводная таблица – раскрывает подробно информацию о продажах, количестве и средней маржинальности по годам и странам, цветовое отображение таблицы соответствует цветовому отображению на карте;
- 5) Текстовый блок: для подведения итогов анализа.

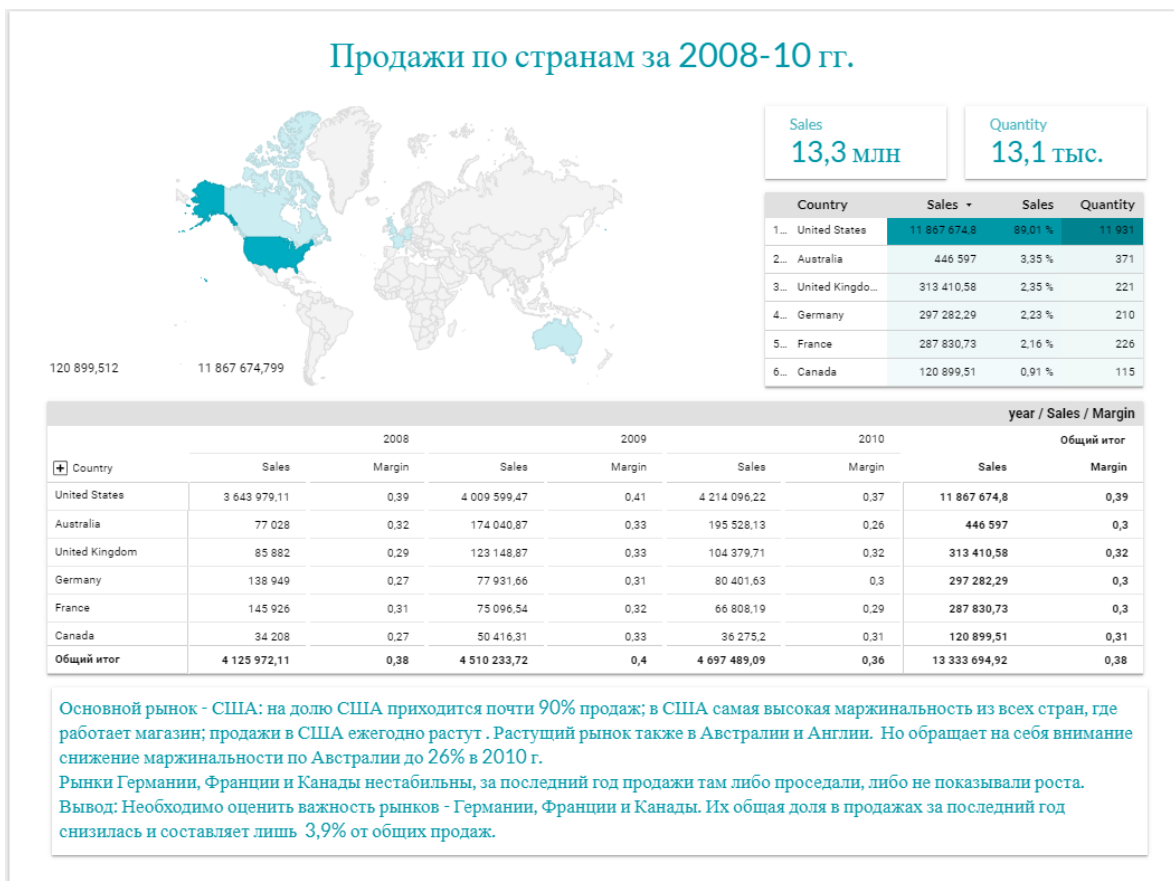
Ссылка на визуализацию: <https://datalens.yandex/tba6wj5wjpwk>



Чтобы ответить на поставленный вопрос в Looker Studio нам будут необходимы следующие элементы:

- 1) Географическая диаграмма – визуально продемонстрирует объем продаж в каждой стране;
- 2) Сводка с комплексными числами – для демонстрации итогов по продажам и количеству;
- 3) Таблица с тепловой картой – как дополнение к Географической диаграмме, которая тоже отражает цветом страны в зависимости от их продаж и дополняет информацию демонстрируя долю продаж каждой страны и количество проданных товаров;
- 4) Сводная таблица: демонстрация продаж и средней величины маржи по периодам;
- 5) Текстовый блок: для подведения итогов анализа.

Ссылка на визуализацию: <https://datastudio.google.com/reporting/697c7eca-05f4-4765-a142-94d6c8a8cd73>.



8. Выводы по работе

№	Задача	Результат																
1	Спрогнозировать продажи на следующий квартал.	<p>Лучшие результаты по анализу временного ряда нашего датасета показала платформа Loginom.</p> <p>Прогноз продаж на последующие 3 месяца:</p> <table><tr><th>Период</th><th>Прогноз</th><th>Нижняя граница</th><th>Верхняя граница</th></tr><tr><td>Январь 2011</td><td>418 087</td><td>282 921</td><td>553 254</td></tr><tr><td>Февраль 2011</td><td>401 735</td><td>262 439</td><td>541 031</td></tr><tr><td>Март 2011</td><td>449 056</td><td>308 466</td><td>589 646</td></tr></table>	Период	Прогноз	Нижняя граница	Верхняя граница	Январь 2011	418 087	282 921	553 254	Февраль 2011	401 735	262 439	541 031	Март 2011	449 056	308 466	589 646
Период	Прогноз	Нижняя граница	Верхняя граница															
Январь 2011	418 087	282 921	553 254															
Февраль 2011	401 735	262 439	541 031															
Март 2011	449 056	308 466	589 646															
2	Построить модель, которая будет предсказывать вероятность возврата проданного товара.	С небольшим отрывом победила модель от Pycaret «Gradient Boosting Classifier» (метрики качества AUC = 0.8129 (F1 =0,4421)), при этом модель Логистической регрессии от Loginom отработала лучше аналога в Pycaret.																
3	Провести анализ клиентов.	<p>Вывод: «идеальных» клиентов, которые одновременно относятся к группе «А» по ABC-анализу и принадлежат одному из относительно положительных классов RFM-анализа, ВСЕГО 143 ЧЕЛОВЕКА из 9 586. Причина в том, что основная масса клиентов приходит в магазин один раз и больше не возвращается.</p> <p>Список «хороших» клиентов, которым можно предложить особые условия обслуживания в первую очередь (143 человека), в Приложении №1.</p> <p>Портрет «идеального» клиента: в среднем ему 45 лет, доход около 60 тыс.ден.ед., имеет 1 или 2 машины, средний чек 8670 ден.ед, покупает 2 или 3 товара.</p>																
4	Провести анализ ассортимента.	<p>Более половины товаров принадлежат группе CX – это товары низкой прибыльностью, но со стабильным спросом (251 штуки). Товары двух наилучших групп AX и VX в общей сложности 26% от активной номенклатуры. На пред заказ необходимо перевести 62 товара.</p> <p>Анализ ассортимента с рекомендациями по складским запасам будет приложен в Приложении №2.</p>																
5	Дашборд «Продажи по странам»	<p>Основной рынок - США: на долю США приходится почти 90% продаж; в США самая высокая маржинальность из всех стран, где работает магазин; продажи в США ежегодно растут. Растущий рынок также в Австралии и Англии. Но обращает на себя внимание снижение маржинальности по Австралии до 26% в 2010 г.</p> <p>Рынки Германии, Франции и Канады нестабильны, за последний год продажи там либо проседали, либо не показывали роста.</p> <p>Вывод: Необходимо оценить важность рынков - Германии, Франции и Канады. Их общая доля в продажах за последний год снизилась и составляет лишь 3,9% от общих продаж.</p>																