

Федеральное государственное образовательное бюджетное
учреждение высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)**

**Институт развития профессиональных
компетенций и квалификаций**

ИТОГОВАЯ РАБОТА

Группа обучения	«ТГУ_АД_2_2022»
Срок обучения	«05.06.2022-»
«ИВАНОВА ИВАННА ИВАНОВНА»	
Номер Кейса	«10»
Название Датасета	«О продажах игровых вертолётов и аэропланов в США»

Москва 2022 г.

Ссылка на файл в Loginom:

https://drive.google.com/file/d/1-u8sGEWTQQRwwT8BJZ4b8gIXBXPii_H/view?usp=sharing

Ссылка на файл в Colaboratory:

<https://colab.research.google.com/drive/1sSerepBeeejNTEvj7BUp6yNOmiTXYVwG?usp=sharing>

Ссылка на работу в Tableau:

https://public.tableau.com/views/Fomina/1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Ссылка на очищенные и подготовленные данные:

<https://docs.google.com/spreadsheets/d/17NyOLQnG-HDsKE1kbD-zgHJOObSUDQIs/edit?usp=sharing&ouid=108518408270365171828&rtpof=true&sd=true>

Описание кейса

Набор данных кейса 10 содержит информацию о продажах игровых вертолётов и аэропланов в США в период 2017-2019 годов. По каждому заказу (номер заказа - уникальное значение) данные содержат ID и название продукта, кол-во товаров в заказе, закупочную цену и цену продажи, величину скидки, а также промокод на скидку (при наличии скидки), дату заказа (день, месяц, год), дату отправки заказа (день, месяц, год), ID, буквенный код и название штата покупателя, ID и название региона покупателя, данные по характеристикам товара (артикул, кол-во каналов, тип товара, тип комплектации, группа товара). Кроме того, исходные данные после обработки данных в программе Loginom были обогащены новыми признаками: продажи, прибыль, ABC-XYZ характеристики товара.

В Loginom производилась очистка и подготовка данных, расчёт новых признаков (созданы новые столбцы Sales, Profit, ABC-XYZ) и других необходимых показателей.

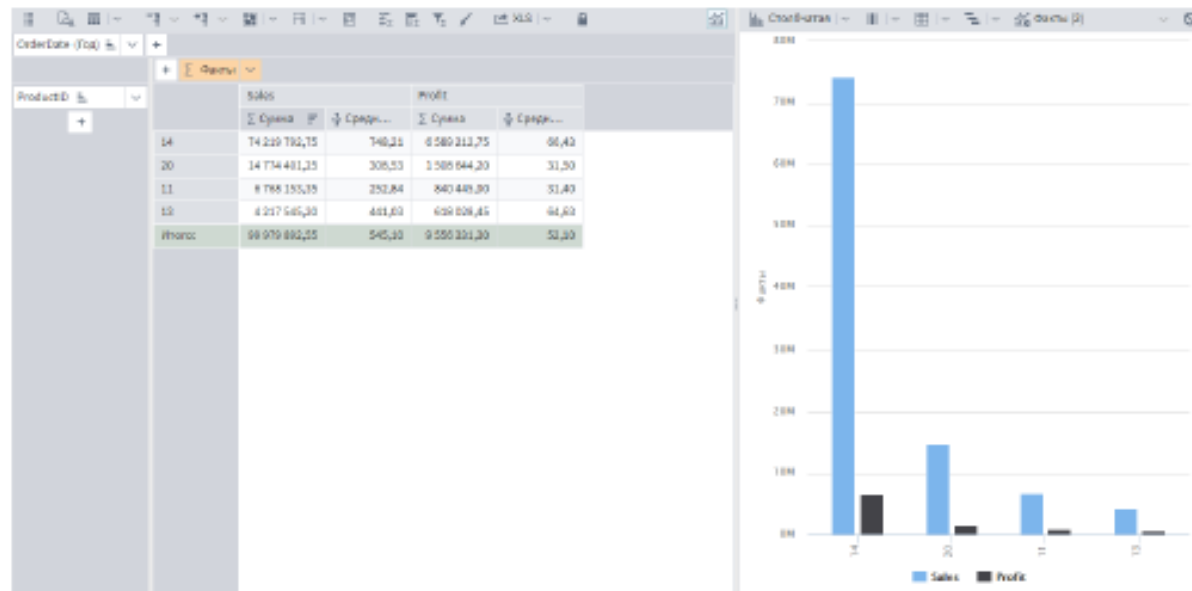
Результаты проведения ABC-XYZ анализа: товаров по классификации ABC - XYZ получилось всего 5 категорий. Это AZ товары (14, 20 и 11 товары) - наиболее часто покупаемые (формирующие 80% выручки и имеющие нестабильный спрос), BZ (товары 4, 16, 10, 18, 13 и 12) - вторые по количеству заказов (формирующие 15 % выручки и имеющие нестабильный спрос), CZ (товары 6, 3, 8, 5, 7, 1, 9, 2, 19) - мало покупаемые (формирующие 5% прибыли и имеющие нестабильный спрос). Почти не покупаемые товары - это товар CY (17 товар), формирующий 5% прибыли и имеющий условно стабильный спрос; и CX (15 товар), формирующий 5% прибыли и имеющий стабильный спрос).

По результатам анализа получаем, что товары, приносящие 80% прибыли, пользуются нестабильным спросом. В этом есть опасность нестабильности прибыли. Возможно, решением этой проблемы будет стимулирование спроса на данные товары (система скидок, например). С другой стороны, покупка дорогостоящего товара (это

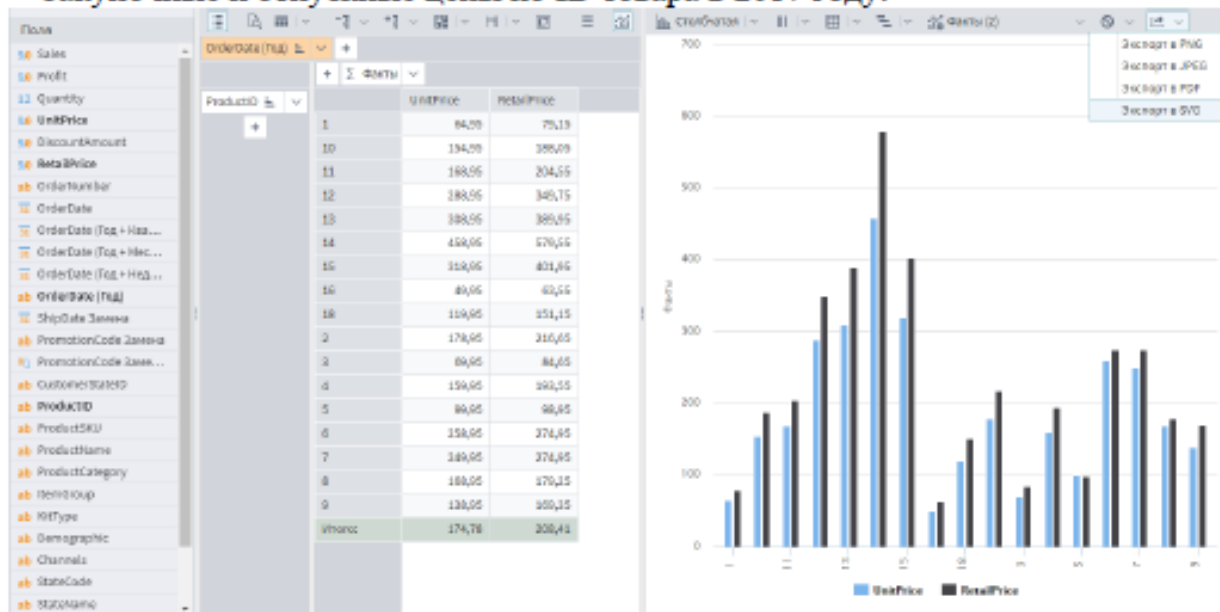
в нашем случае так) вряд ли будет частой у одного и того же покупателя, т.к. товар длительного пользования. Возможно привлечение новых клиентов для стабилизации спроса (например, реклама в те периоды, в которые продажи наиболее ходовых товаров падают).

В Logiplot также производился исследовательский анализ данных. Некоторые интересные и значимые показатели графики приведены ниже:

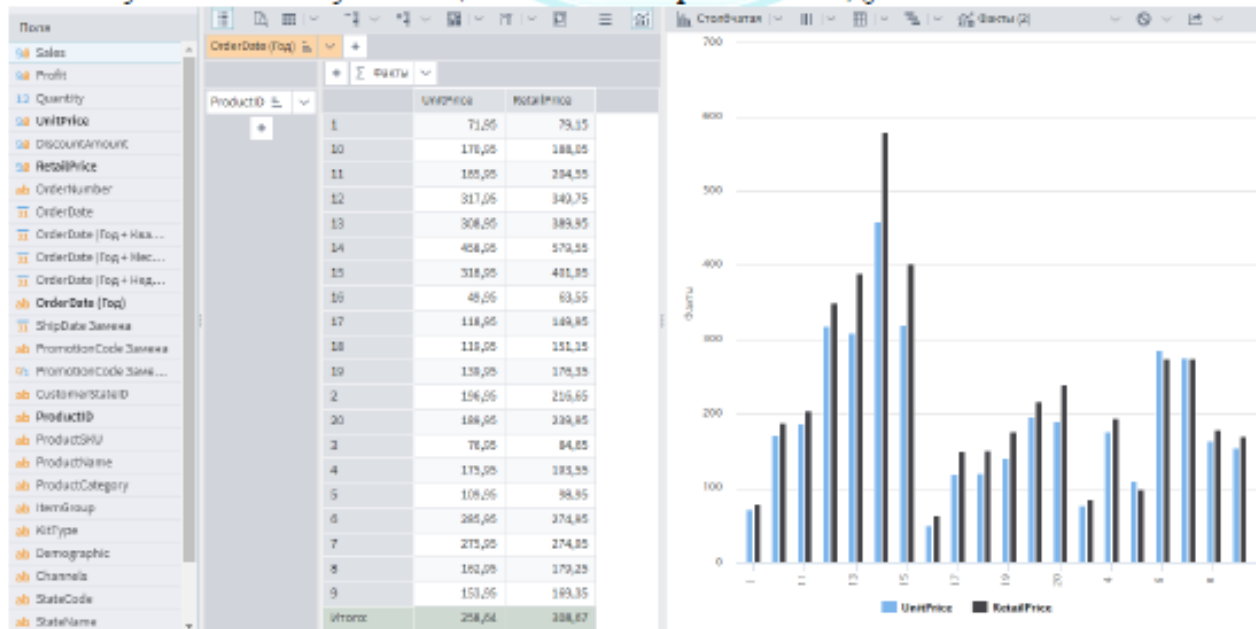
80% продаж по ID продукта за всё время:



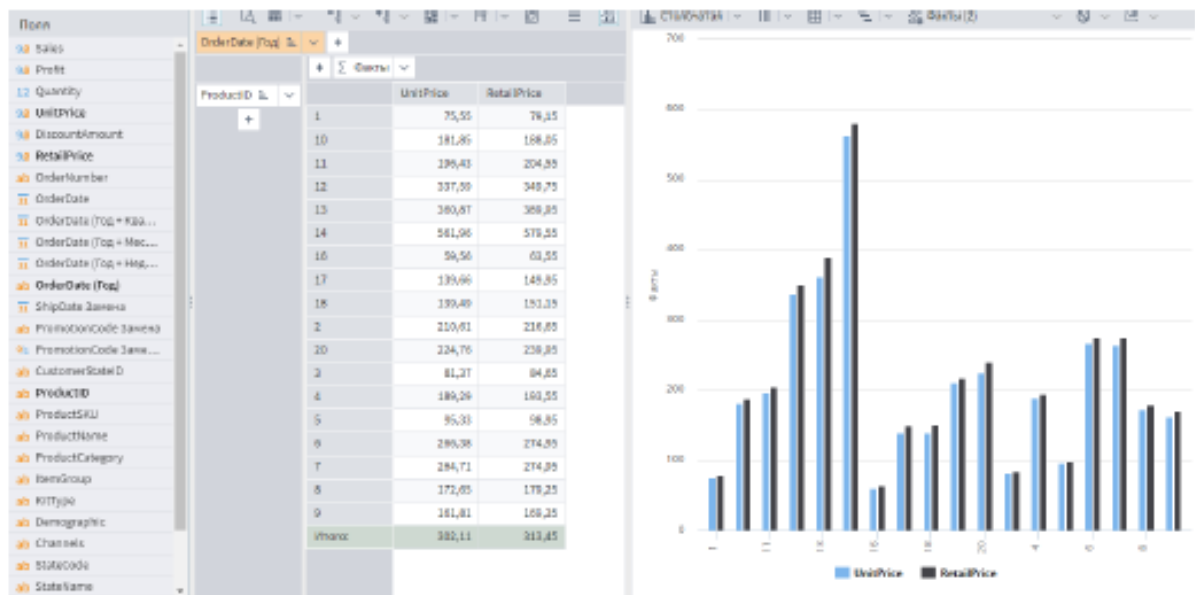
Закупочные и отпускные цены по ID товара в 2017 году:



Закупочные и отпускные цены по ID товара в 2018 году:



Закупочные и отпускные цены по ID товара в 2019 году:



Из трёх приведённых графиков видно, что причина резкого падения прибыли в том, что закупочные цены возросли, а отпускные остались прежними. В 2019 году отпускные цены практически сравнялись с закупочными. Необходимо пересмотреть ценообразование и ценовую политику. Более подробно анализ данных произведён в Colaboratory, ссылка на файл приведена на первой странице.

Основные результаты EDA:

- Количество товаров в заказах от 1 до 3. Наибольшее количество заказов - покупка 1 товара.
- Наиболее популярный товар в заказе - это товар с ID 14, затем 20 и 4. Это тройка лидеров. Далее следует товар 11, 6, 16, 10. Наименьшей популярностью пользуются товары 15, 9, 17.
- Наиболее часто покупаемый товар 14 имеет наивысшую среднюю закупочную цену (522,27) и отпускную цену (579,55). Второй по популярности товар 20 имеет среднюю закупочную цену 213,84 и отпускную 239,95 (восьмой по величине закупочной цены). Третий по популярности товар 4 - средняя закупочная цена 184,05, отпускная 193,55. Следующий по популярности товар 11 в среднем закупается за 178,62 и продаётся за 204,55 (10-й по величине закупочной цены). И пятый по популярности товар 6 имеет среднюю закупочную цену 271,73 и отпускную 274,75 (5-й по величине закупочной цены). Вывод - наиболее часто покупают самый дорогой товар. А самый дешёвый товар с ID 16 - шестой по популярности.
- Всего 6 категорий товара. Наиболее часто покупаемая категория - это Collective pitch, затем Trainer и Warbrid, наименее редко заказываемая - Fixed pitch.
- Групп товаров всего 2, приблизительно равные по количеству. Это Helicopter (чуть больше по количеству в заказах) и Airplane.
- Чаще покупают товар в комплектации RTF, чем KIT. Но различие в величине незначительное.
- Чаще заказывают товары с количеством каналов 6, затем 4, далее 5 и 3.
- Подавляющее количество заказов были оформлены без скидки, т.е. промокод на скидку не применялся.
- Данные представлены за 3 года – с 2017 по 2019 года. Больше всего заказов сделано в 2019 году, наименьшее количество в 2017 году.
- Наиболее часто покупают товар в категории Professional, затем Intermediate, далее Advanced Demographic. Это подтверждает то, что чаще заказывают самый дорогой товар с ID 14, являющийся товаром для профессионального использования.
- Чаще всего товары заказывают покупатели из Калифорнии, Флориды и Нью Йорка, меньше всего заказов из Вайоминга.
- Наибольшее число заказов из региона Southern, затем Midwest, далее практически поровну из Northeast и Southwest регионов.
- Наблюдается очень негативная тенденция - при увеличении продаж по годам прибыль изначально растёт, но затем падает! Наибольшая прибыль в период с апреля 2018 г. по апрель 2019 года при постоянном увеличении объёма продаж. А

отрицательной прибылью стала в июле 2019 года, наиболее убыточные значения в декабре 2019 года.

- Для решения вышеприведённой проблемы была проанализирована общая тенденция по соотношению цен закупки и продажи по товарам. Было выявлено, что причина такого резкого падения прибыли – закупочные цены товаров достаточно сильно выросли с 2017 года, а значения отпускных цен практически не изменились. В 2019 г., особенно в последнем квартале года, эти показатели практически сравнялись (средняя закупочная цена 302, а средняя цена реализации 313,45). Это является основной причиной такого падения прибыли! Рекомендация - либо найти возможности снизить закупочные цены (скидки, другой поставщик и пр.), либо придётся повышать отпускные цены на товары.

Следующий этап – машинное обучение.

И в Colaboratory, и в Logiном была произведена линейная регрессия, искомым признаком регрессии – Sales (продажи).

Результаты линейной регрессии в Logiном приведены ниже:

Модель	Фактический результат	Атрибут	Коэффициент	Стандартная ошибка	T-статистика	P-значение	Нижняя граница
Показатель	Значение						
Константа	Волонка	ab_abc_002	9,815654	6,879855	134,813767	0,000000	
Латентная функция производства	-1,808175,802074	ab_CV	4,360559	0,800000	5,450694	0,000000	
Коэффициент детерминации	0,906843	ab_C2	-3,767749	0,387573	-9,720652	1,000000e-09	
Коэффициент детерминации (корр.)	0,906813	ab_CX	-10,833981	0,839432	-12,906131	0,000000	
Стандартное отклонение	96,837332	ab_C3	334,882523	71,117,879833	0,004388	0,999999	
Число степеней свободы ошибок	382,050,09	ab_константа	25,385235	2,251,793419	0,013048	0,989989	
Число степеней свободы модели	97,800000	ab_Product00					
R-статистика	28,294,338273	ab_1	0,570159	826,828924	0,013985	0,990004	
R-квадрат модели	0,951136-18	ab_17	0,294982	895,219041	0,013191	0,987880	
Критерий Акаике	11,807973	ab_18	8,923972	897,820534	0,009042	0,992088	
Критерий Акаике (корр.)	11,807978	ab_8	7,529739	823,138708	0,008157	0,993492	
Критерий Байеса	11,877118	ab_5	5,042735	883,305910	0,005735	0,995440	
Критерий Акаике-Нулика	11,808073	ab_10	4,223187	1,016,344904	0,004158	0,996584	
		ab_12	1,886135	888,343852	0,002123	0,998306	
		ab_19	1,605530	438,123377	0,003985	0,997076	
		ab_7	0,238087	850,883096	0,002277	0,999779	
		ab_9	-0,577150	727,738208	-0,000782	0,999376	
		ab_2	-3,725982	724,305529	-0,003783	0,998897	
		ab_6	-5,173630	1,183,828450	-0,004585	0,998254	
		ab_16	-0,022098	1,185,376558	-0,000448	0,998553	
		ab_4	-6,713275	1,153,788453	-0,005267	0,985718	
		ab_3	-6,715712	885,818624	-0,006969	0,986884	
		ab_11	-7,778686	1,138,532703	-0,006835	0,984554	

Коэффициенты регрессионной модели:


#	Имя входных полей	Метод входных полей	Уникальные значения	Sales Коэффициенты P1	Sales Cтг, откл.	Sales T-статистика	Sales Значимость	Sales P-коэф.
1	Quantity	Quantity	234,811222	71117,57263	0,804568230881	0,9862308068		
2	«Константа»		29,28925405	2251,709419	0,01264881258	0,9895849893		
3	CustomerStateID	CustomerStateID	28	18,48936286	5418,242182	0,803043264886	0,9875717987	
4	ABC_XYZ	ABC_XYZ	CV	8,811853884	8,07866459367	134,4127673	0	
5	ProductID	ProductID	1	8,578155081	820,5269342	0,01168342687	0,9800941525	
6	ProductID	ProductID	17	8,254562213	609,219041	0,01519085173	0,9878798215	
7	ProductID	ProductID	18	8,823975035	857,8205142	0,008941888886	0,9820877215	
8	ProductID	ProductID	8	7,528758893	823,1287645	0,008156670293	0,9824918048	
9	CustomerStateID	CustomerStateID	8	6,798964278	5726,112679	0,801187264755	0,9895381135	
10	CustomerStateID	CustomerStateID	50	6,538880881	3008,444044	0,802174390793	0,9882858255	
11	CustomerStateID	CustomerStateID	49	6,452455089	5011,569101	0,802081487128	0,9891261082	
12	ABC_XYZ	ABC_XYZ	C2	6,39858685	0	0	0	
13	CustomerStateID	CustomerStateID	41	6,23388284	4234,264382	0,803471675885	0,9886255277	
14	CustomerStateID	CustomerStateID	48	5,343123073	5287,550571	0,8080782252074	0,9892154825	
15	CustomerStateID	CustomerStateID	27	5,84206925	6075,452486	0,808362682851	0,9892315323	
16	ProductID	ProductID	5	5,843738237	882,3059182	0,805715485813	0,9854397053	
17	CustomerStateID	CustomerStateID	21	4,853808832	5881,081583	0,8088421630316	0,9892380439	
18	Channels	Channels	3	4,458136201	39256,45885	0,8082382176706	0,898818289	
19	Demographic	Demographic	Beginner	4,459339781	3038,58518	0,804072626395	0,9886753293	
20	CustomerStateID	CustomerStateID	29	4,258845275	4817,208075	0,8088842675858	0,9892844285	
21	ProductID	ProductID	10	4,223187089	3035,244684	0,80415678345	0,988842582	
22	RegionID	RegionID	2	4,08852733	34311,80677	0,8083893287731	0,9897720773	
23	CustomerStateID	CustomerStateID	26	1,754307487	4888,479449	0,806787680853	0,9892872284	
24	CustomerStateID	CustomerStateID	16	1,873162386	8558,569882	0,805481627386	0,9895828083	

Статистические характеристики по регрессии:

#	Метка	Вид	Гистограмма	Диаграмма...	Минимум	Макс...	Среднее	Стандартное отклонение	MP	Метка	Доля	Кал...	%
1	00 SalesPherson	0			8,81186...	1312,8...	375,08...	286,45341791873	4	130,5	208,8	58287	24
2	00 ProductID	0	Число значений - 28	Нормальное	1	2	1,8940...	0,478842315881207	11	617,4	886,8	73522	29
3	00 ABC_XYZ	0		Нормальное	2	2	2	0	5	250,8	208,8	69880	17
4	00 Sales	0			37,85	1738,85	375,125...	395,844773029308	2	-5,0	68,8	38745	39
5	12 Quantity	0			1	5	1	0,513922997413933	9	478,2	947,8	23447	5
6	00 CustomerStateID	0			48,35	579,55	272,235...	384,893882985707	7	335,3	408,7	18879	5
7	00 DISCOUNT...	0			0	434,85	5,48438...	32,1291599182897	18	985,3	1094,7	15427	4
8	00 Model Fit R	0			63,35	579,55	299,452...	380,125947175887	3	85,9	139,3	14989	4
9	00 OrderDate	0			03.01.2017	31.12...	03.01.20...	282,8932178125861	6	289,8	338,1	8239	2
10	00 OrderDate (год)	0		Нормальное	4	4	4	0	10	587,8	617,4	6368	2
									8	486,7	478,2	8287	2
									20	1242,3	11...	5882	2
									13	756,5	826,8	4229	1
									15	885,8	905,1	6089	1

Линейная регрессия, произведённая в Colab, также показала очень хорошие результаты. Так, R2 имеет очень высокое значение 0,9057. Наиболее значимый признак Quantity, затем OrderDate_Y_1, затем признаки KitType и ABC_XYZ. Если сравнивать с моделью, построенной в Logiom, то первый по значимости признак совпадает. В модели линейной регрессии Logiom коэффициент детерминации равен 0,9, а стандартное отклонение 96,04%.

При использовании модели RandomForest даже всего только с 2 деревьями, с максимальной глубиной 5 и минимальным количеством примеров в листе 5 модель на учебном множестве переобучается. На тестовом множестве модель показывает крайне высокую точность 99,47% (0,9947), что выше, чем алгоритм линейной регрессии. Но в данном случае имеет место переобучение. Так как точность при линейной регрессии также очень высока, но алгоритм проще и прозрачнее, лучше пользоваться им.



По результатам применения градиентного бустинга модели GradientBoostingRegressor, LGBMRegressor, HistGradientBoostingRegressor показали себя хуже, чем модели линейной регрессии. Так, при выборе количестве деревьев (итераций) 5 R квадрат на тестовом множестве колеблется от 63,03% (GradientBoosting) до 65,8% в случае LGBM и HistGradientBoosting. Это значительно ниже результатов линейной регрессии и метода RandomForest. При количестве деревьев 10 результат значительно лучше. R2 уже приближается к 90%. Так, для GradientBoosting это значение 86,01%, для LGBM и HistGradientBoosting 87,79%. Это всё равно ниже значений первых двух алгоритмов, данные модели значительно более сложные, требуют больше времени и ресурсов для работы. Поэтому целесообразно остановиться в нашем случае на первой модели - Линейной регрессии.

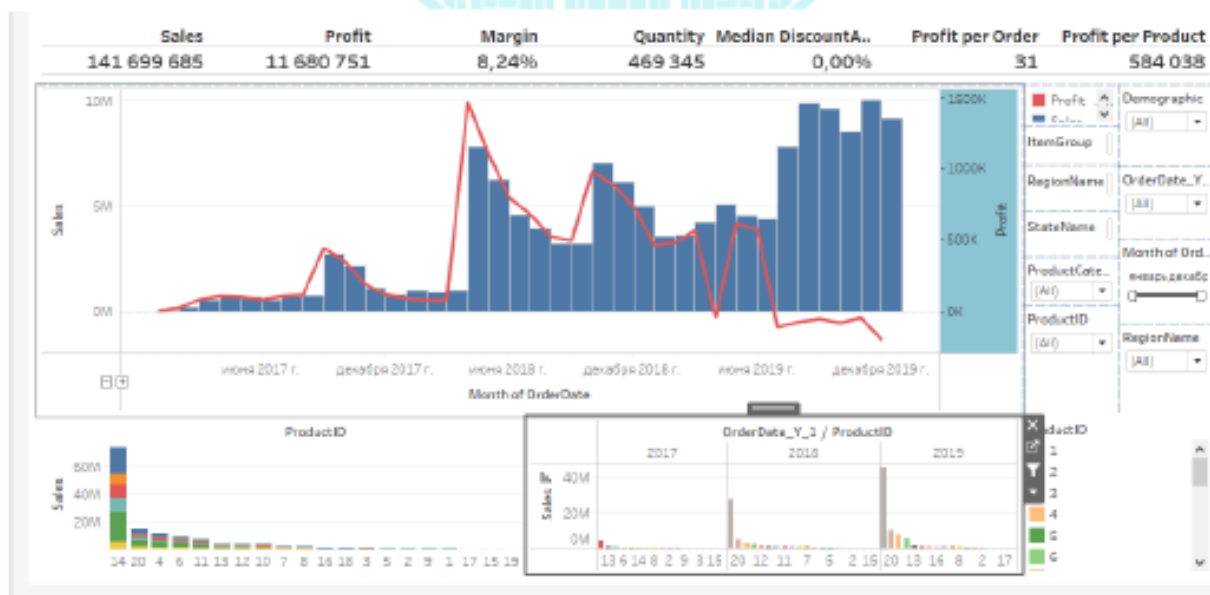
Показатели модели линейной регрессии, построенной в Loginom: стандартное отклонение модели получилось 96,04%, коэффициент детерминации 0,9. Значения высокие, сопоставимы с моделью линейной регрессии, построенной в Colab. Но в Loginom обработка, анализ данных осуществляются намного проще, без написания кода. Отсюда делаю вывод, что наиболее оптимально обрабатывать данные, строить модель в Loginom. А обработанные в Loginom данные удобно использовать в таких продуктах, как Tableau или Power BI для дальнейшей визуализации и построения дашбордов. Однако в данной программе нет возможности использовать разные предсказательные модели, что лишает пользователя возможности сравнить успешность различных моделей.

Следующим этапом работы было построение дашбордов в Tableau.

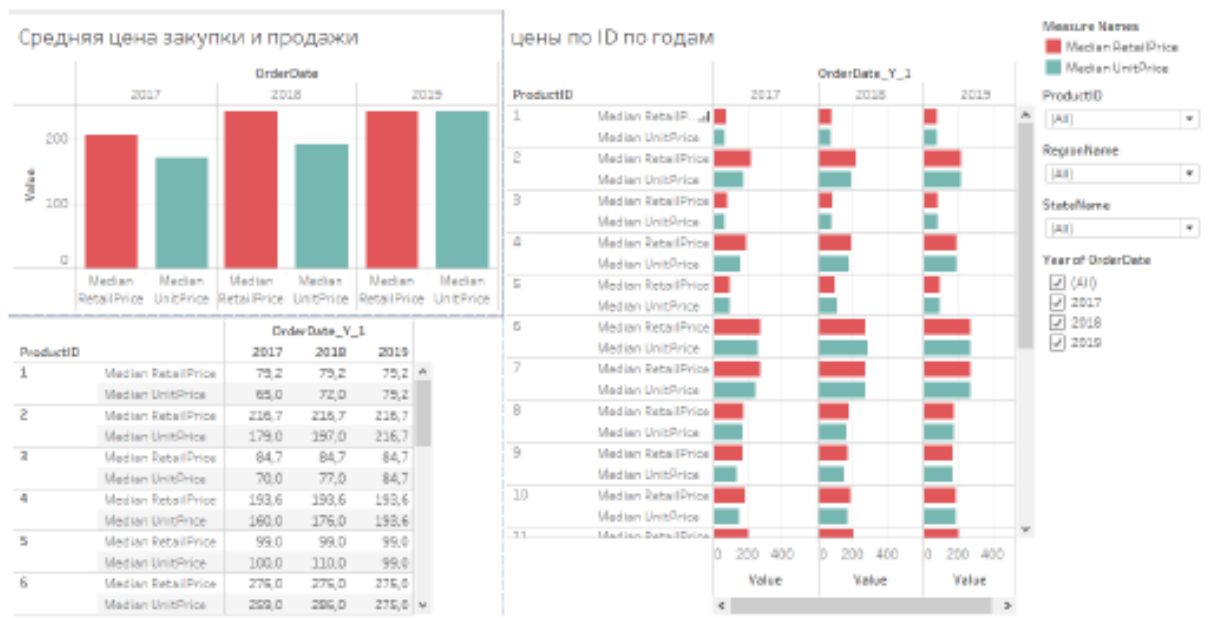
Ниже приведены построенные дашборды.

Первый дашборд показывает основные значения KPI (продажи, прибыль, маржа, количество заказов, средняя скидка, средняя прибыль на заказ и прибыль по продукту).

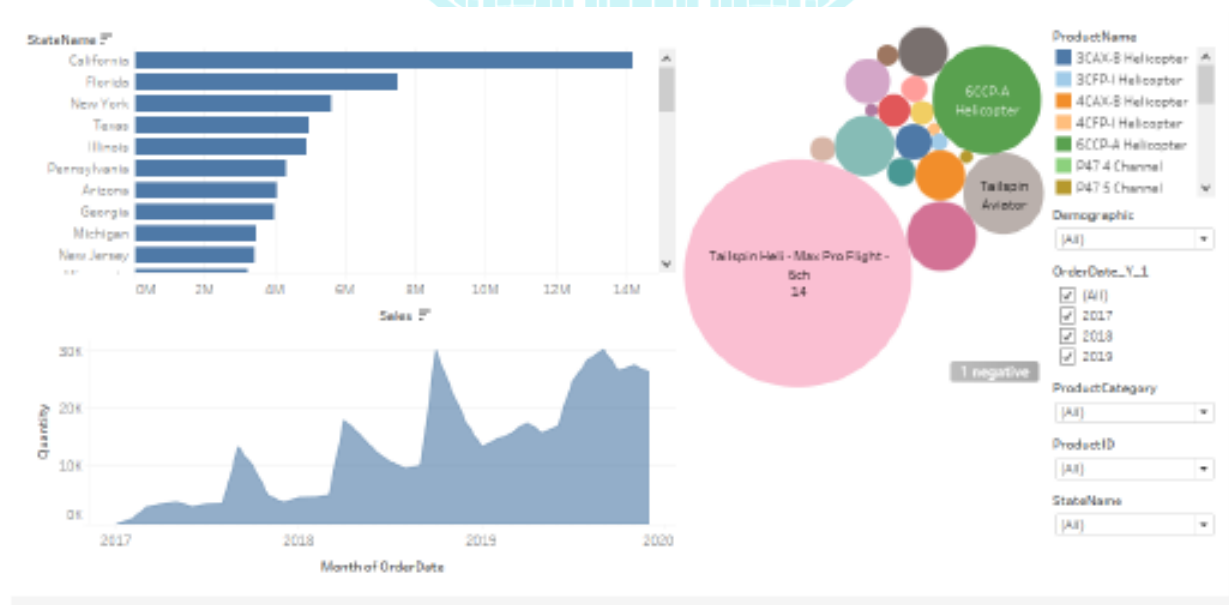
Отображена основная тенденция увеличения объёмов продаж с 2017 по 2019 годы и увеличение, а затем падение прибыли. Также проранжированы товары по величине продаж, в том числе по цветовой кодировке видно количество продаж каждого товара по регионам. И проиллюстрированы наиболее популярные товары по каждому году.



Следующий дашборд визуализирует средние цены закупки и продажи для каждого товара и в целом по товарам по каждому году. Также можно пользоваться дополнительной фильтрацией по регионам, штатам, кроме фильтрации по годам, ID товара, цене закупки или продажи.



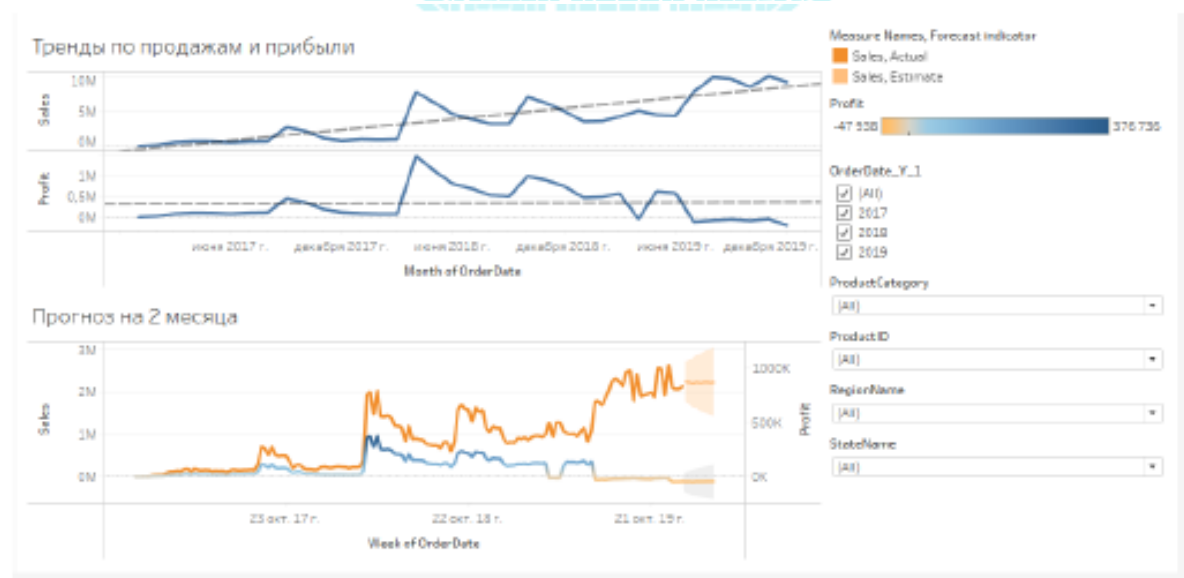
Третий дашборд показывает величину продаж по штатам, динамику продаж по количеству продаж за весь период, а также объём продаж по товарам (их ID и названиям). Также установлены фильтры по названию товара, году заказа, категории продукта, ID продукта, названию штата.



Далее построены визуализации по ABC-XYZ анализу товара. Видно, какие товары входят в какую категорию, величины объёма продаж и прибыли по каждому товару.

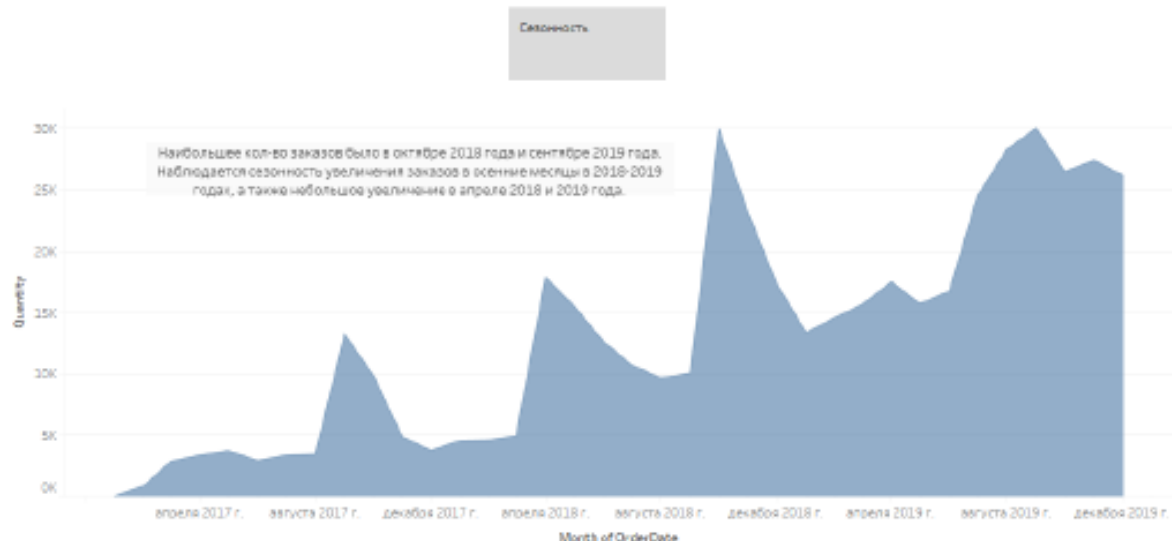


Следующая визуализация – это тренды по продажам и прибыли, а также прогноз по продажам и прибыли, построенный на 2 месяца вперёд. Дополнительно можно производить фильтрацию в том числе по ID, категории продукта, названию региона, названию штата.



Кроме дашбордов, было сделано несколько историй – заметок. Все они приведены ниже:

Сезонность



Динамика продаж и прибыли

Динамика продаж и прибыли

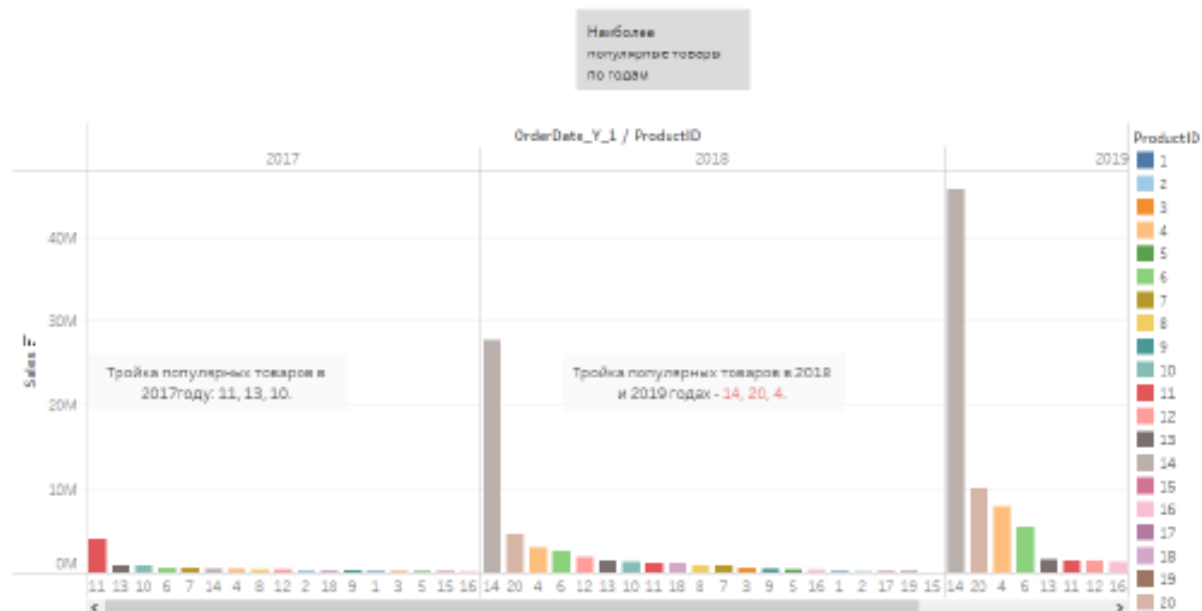


Число продаж по ID товара в регионах

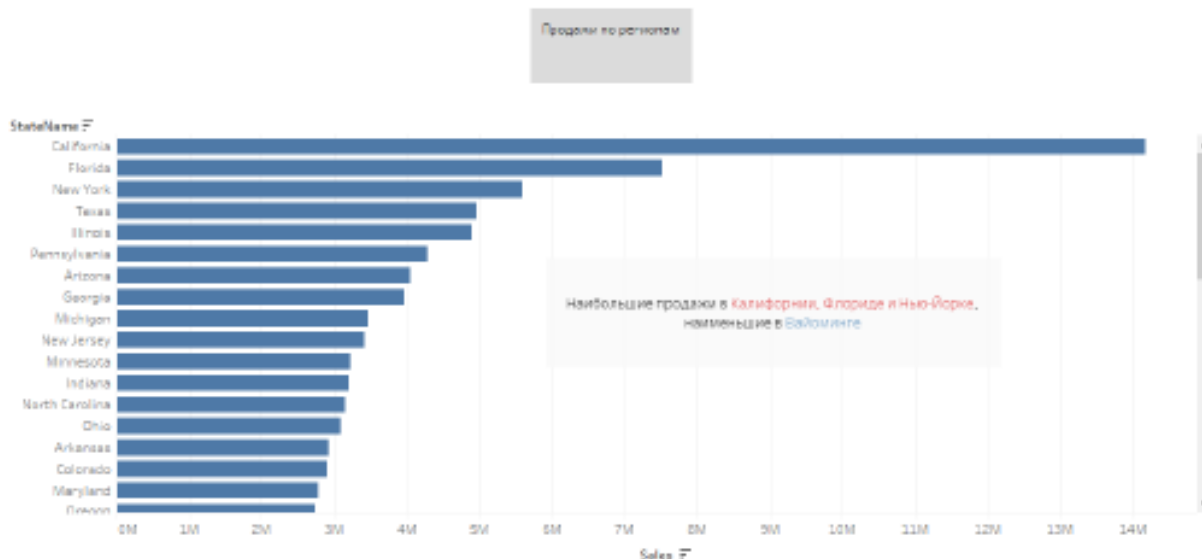
Число продаж по ID товара в регионах



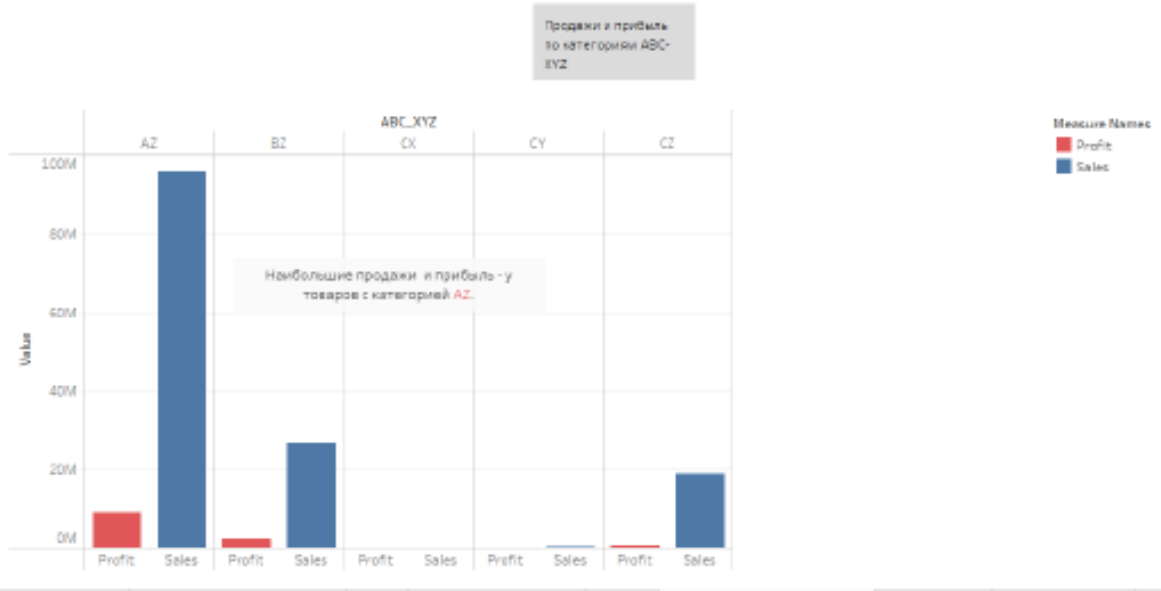
Наиболее популярные товары по годам



Продажи по регионам



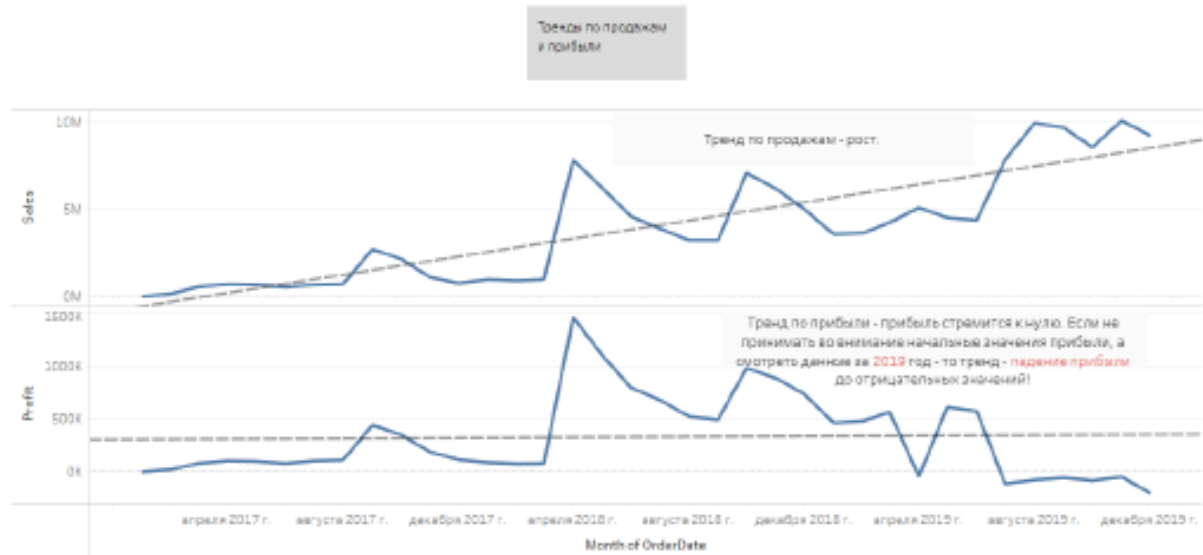
Продажи по категориям ABC-XYZ



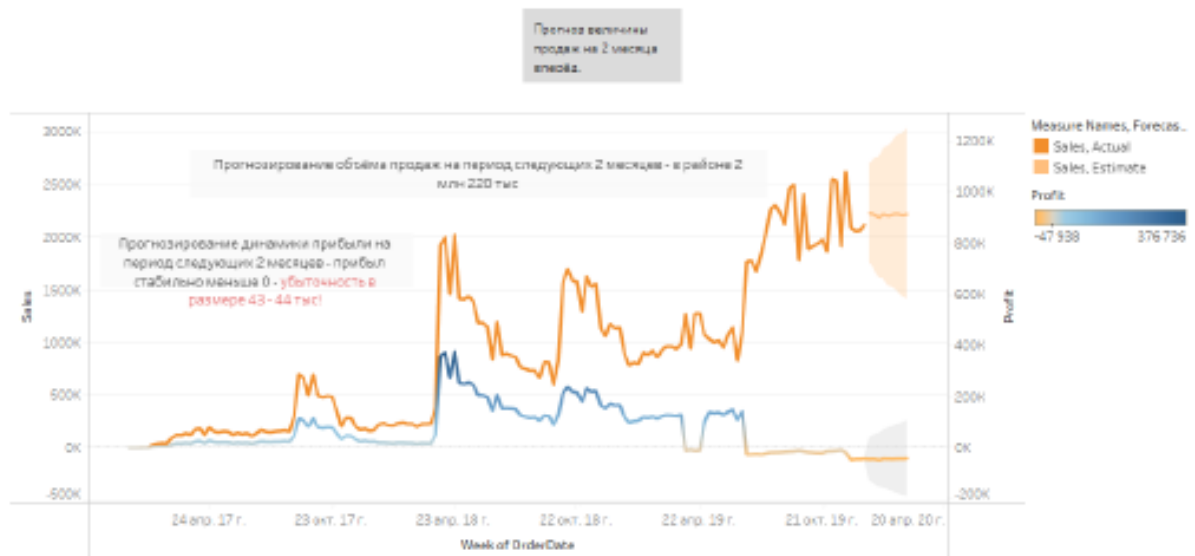
Цена закупки и продажи по годам



Тренды



Прогноз



Построенные в Tableau визуализации также были встроены в блокнот Colab.