Федеральное государственное образовательное бюджетное учреждение высшего образования

«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Институт развития профессиональных компетенций и квалификаций

ИТОГОВАЯ РАБОТА

Группа обучения	**«2_АД_СЗ_23»**
Срок обучения	**«11.08.2023-31.08.2023»**
«НАУМОВА ЕЛЕНА АНАТОЛЬЕВНА»	
Номер Кейса	**«40»**
Название Датасета	**«Продажи радиоуправляемых самолетов и вертолетов в США за 2017-2019 гг.»**

Москва 2023 г.

Ссылка на файл в Loginom:

https://drive.google.com/file/d/14iaFVRG6bcDBC4BF7CmUBC6729cQReq1/view?usp=sharing

Ссылка на файл в Google Colab:
https://colab.research.google.com/drive/1jcCwsegF47nlQRpdp-hVI1SpfPfcKgoz?usp=sharing

Ссылка на работу в PowerBI:
https://drive.google.com/file/d/1GGSIrjIlCY9EkCj5Sch2KoJXxbz9tGOk/view?usp=sharing

Ссылка на очищенные и подготовленные данные:
https://docs.google.com/spreadsheets/d/1fjIxi-MuQw5omLqvxKrlXofsMBvdUQ6p/edit?usp=drive_link&ouid=111141607451877612723&rtpof=true&sd=true

1. Описание кейса.

Датасет предоставляет информацию о продажах радиоуправляемых самолетов и вертолетов в США с 01.01.2017 по 31.12.2019 гг.

Данные хранятся в excel-файле на 4 листах:

- Лист 1 – Product – содержит 9 столбцов и 20 строк, содержит список и описание товара:

Product ID - идентификатор товара;

Product SKU - артикул товара;

Product Name - название товара;

Product Category - категория товара;

Item Group - группа товаров;

Kit Type - тип комплектации;

Channels – количество каналов:

Demographic – уровень сложности;

Retail Price – розничная цена.

- Лист 2 – Region - содержит 2 столбца и 6 строк, содержит список регионов:

Region ID - идентификатор региона;

Region Name - наименование региона.

- Лист 3 – Sales - содержит 9 столбцов и 377741 строк, содержит данные о продажах за 2017-2019 гг.:

Order Number - номер заказа;

Order Date - дата заказа;

Ship Date - дата отправки заказа;

Customer State ID – идентификатор штата отправки;

Product ID - идентификатор продукта;

Quantity – количество товара;

Unit Price – закупочная цена за единицу товара;

Discount Amount – сумма скидки по промокоду;

Promotion Code – промокод.

- Лист 4 – State - содержит 4 столбца и 51 строк, содержит перечень штатов.

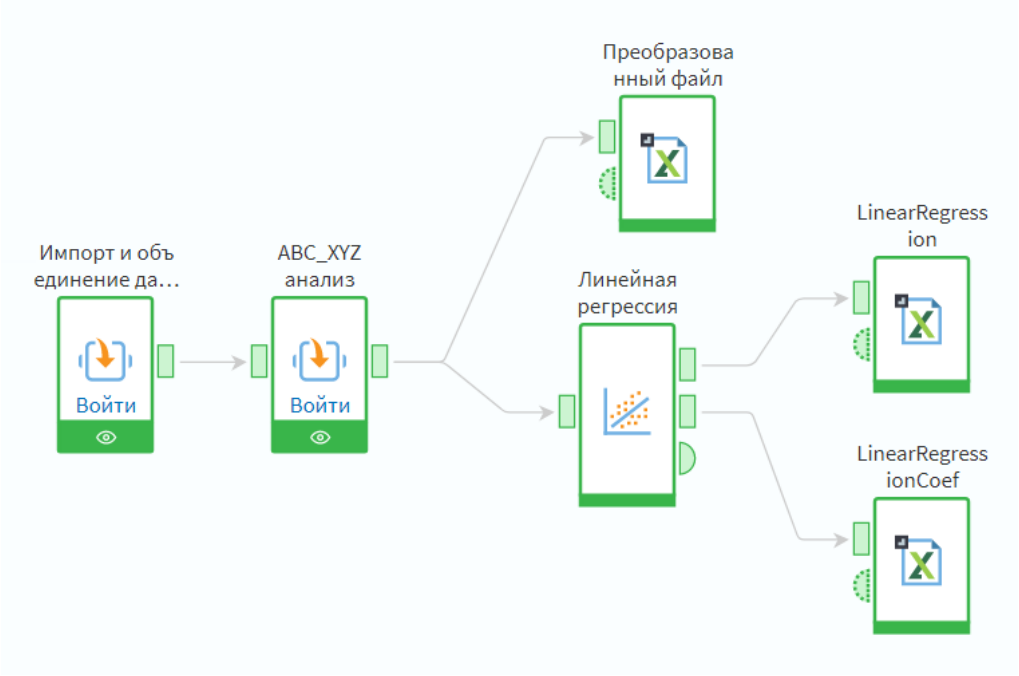
State ID – идентификатор штата;

State Code - буквенный код штата;

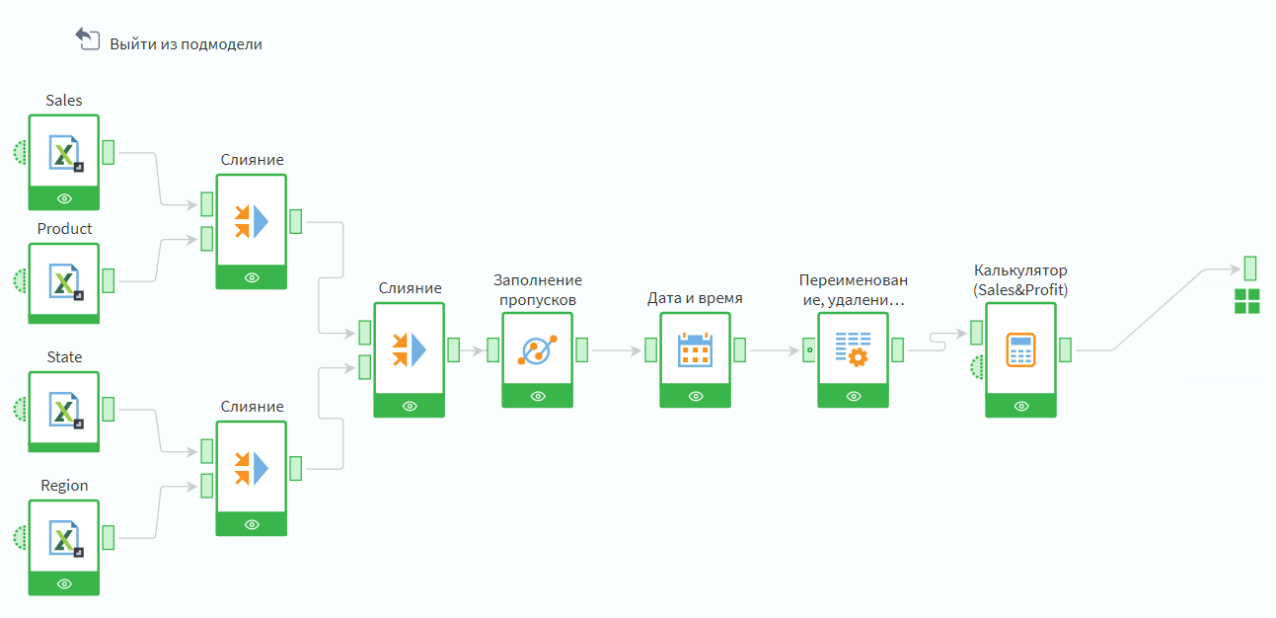
State Name - наименование штата;

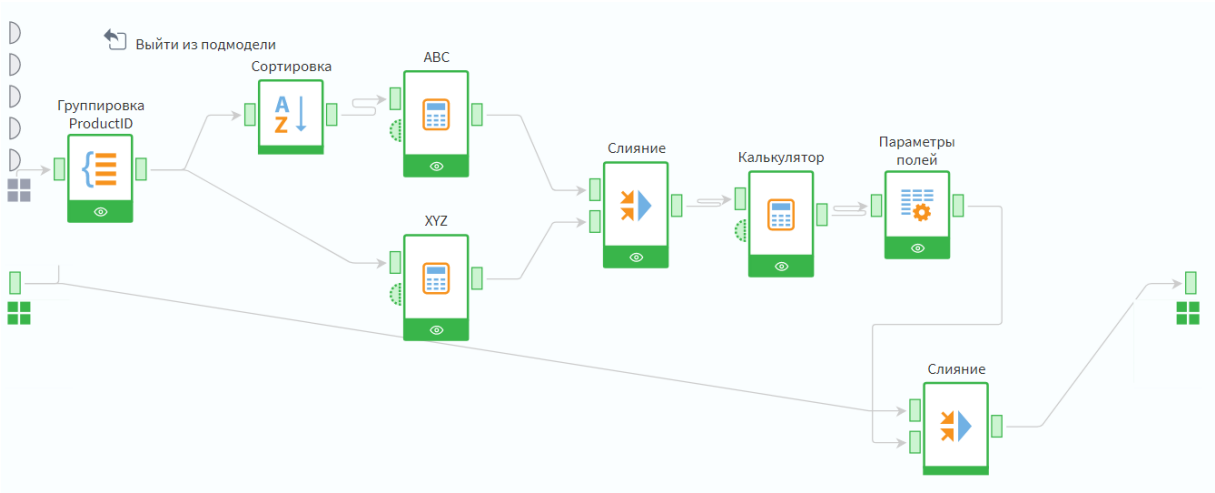
Region ID - идентификатор региона.

С помощью платформы Loginom Community проведены объединение, первичная обработка и первичный анализ данных.



Набор данных дополнен новыми признаками (Sales – сумма продажи по заказу с учетом скидки, Profit – прибыль по заказу, OrderProcessing – срок обработки заказа), а также результатами ABC-XYZ-анализа.





2. Результаты EDA в Google Colab

Value	Count	Frequency (%)
14	99197	26.3%
20	47886	12.7%
4	46923	12.4%
11	26769	7.1%
6	26375	7.0%
16	23713	6.3%
10	16419	4.3%
3	13608	3.6%
8	11908	3.2%
18	9810	2.6%
Other values (10)	55133	14.6%

Наиболее часто покупаемые товары:

- 26,3% - ID 14 (Tailspin Heli - Max Pro Flight - 6c),
- 12,7% - ID 20 (6CCP-A Helicopter),
- 12,4% - ID 4 (Piper Cub 4 Channel),
- 7,1% - ID 11 (Tailspin Aviator Mk2-15).

Эти же товары (ID 14, 20 и 11) составляют категорию A.

Value	Count	Frequency (%)
A	173852	46.0%
B	115909	30.7%
C	87980	23.3%

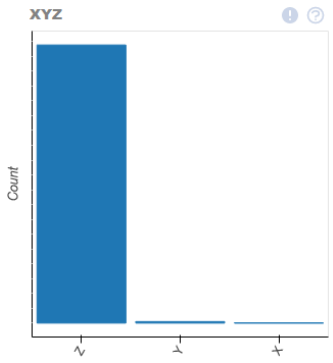
- Категория A - 46,0% - товары с ID 14, 20 и 11 - составляют 80% прибыли.
- Категория B - 30,7% - товары с ID 13, 4, 18, 10, 12, 16.
- Категория C - 23,3% - остальные товары.

При этом прибыль не всегда зависит от продаж (например, у товара с ID 6 продажи высокие, а прибыли нет).

Почти все товары относится к категории Z, т.е. к нерегулярным продажам.

Value	Count	Frequency (%)
Z	375413	99.4%
Y	2080	0.6%
X	248	0.1%

- Категория X - 0,1% - спрос стабильный, регулярный.
- Категория Y - 0,6% - спрос непостоянный, влияние тенденций.
- Категория Z - 99,4% - спрос случайный.



Это, скорее всего, продиктовано особенностями самого товара - не первой необходимости, непотребляемый, т.е. рассчитан на длительное использование, с низкой амортизацией.

Также, если рассмотреть категорию покупателей, которые покупают товар, то можно увидеть, что большинство продаж приходится на товары для профессионалов и уверенных пользователей (Professional - 38,9%, Intermediate - 33,8%).

Value	Count	Frequency (%)
Professional	147083	38.9%
Intermediate	127671	33.8%
Advanced	56267	14.9%
Novice	30513	8.1%
Beginner	16207	4.3%

Таким образом, отсутствуют товары категорий AX, AY, BX, BY:

- Категория AZ - 46,0% - все те же товары с ID, что и в категории A.
- Категория BZ - 30,7% - все те же товары с ID, что и в категории B.
- Категория CZ - 22,7% - большинство остальных товаров.

Value	Count	Frequency (%)
AZ	173852	46.0%
BZ	115909	30.7%
CZ	85652	22.7%
CY	2080	0.6%
CX	248	0.1%

Учитывая специфику товара, отсутствие регулярного постоянного спроса можно компенсировать только привлечением новых клиентов, рекламными кампаниями, увеличением количества товаров в заказе.

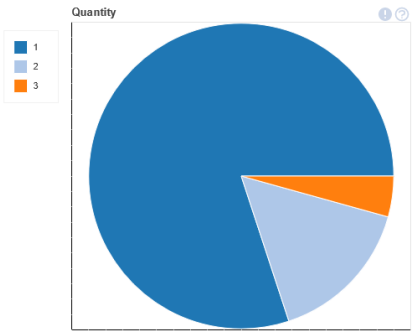
В основном товары реализовывались без применения скидок и промокодов, это 83,7% продаж.

Value	Count	Frequency (%)
0	316033	83.7%

Товары, которые чаще покупали (ID 14, 20, 4, 11, 6) чаще остальных продавались со скидкой. Общая скидка по каждому товару за весь период составляет около 1% и меньше, что отрицательно не сказывается на сумме продаж и прибыли.

Количество товаров в заказе варьируется от 1 до 3 единиц, где:

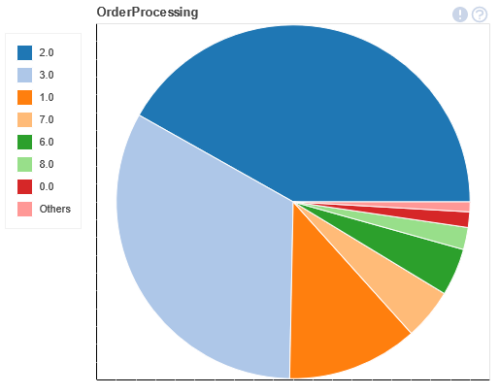
- 1ед - 80,1%. Большинство покупателей приобретают по 1-му товару за раз. Это также продиктовано спецификой товара.
- 2ед - 15,6%.
- 3ед - 4,3%, т.е. крайне редко.



Все значения OrderNumber являются уникальными, т.е. не было заказов, в которых были бы несколько разных позиций товаров. Покупатель чаще всего покупает 1 товар в количестве 1 шт. или же реже от 2 до 3 шт. товаров одного вида.

В среднем заказы до отправки обрабатываются 1-3 дня — это 88%:

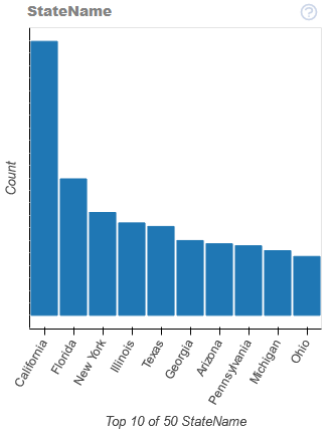
- 2 дня - 41,8%
- 3 дня - 32,8%
- 1 день - 12,0%
- 0 дней - 1,4%



Но есть 12 % заказов, которые обрабатывались 6, 7 или даже 8 дней. По представленным данным невозможно оценить, как это повлияло на продажи, т.к. нет информации по отказам от заказа из-за длительной обработки. Но в любом случае

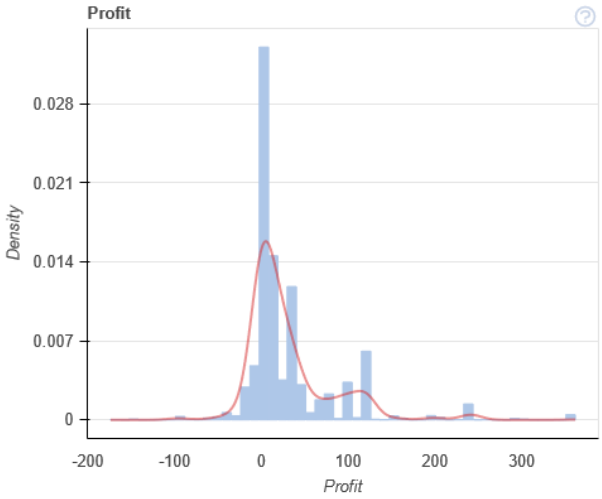
следует контролировать и минимизировать срок обработки заказа более 3 дней, избегать такие длительные задержки.

Большинство товаров доставлялось в Калифорнию, Флориду и Нью-Йорк, наименьшая часть - Вайоминг. Но продажи присутствуют во всех 50 штатах.



У признака Profit имеется 11,63% отрицательных значений. Т.е. в 11,63% случаев продажи производились в убыток.

Negative (%)	11.6%

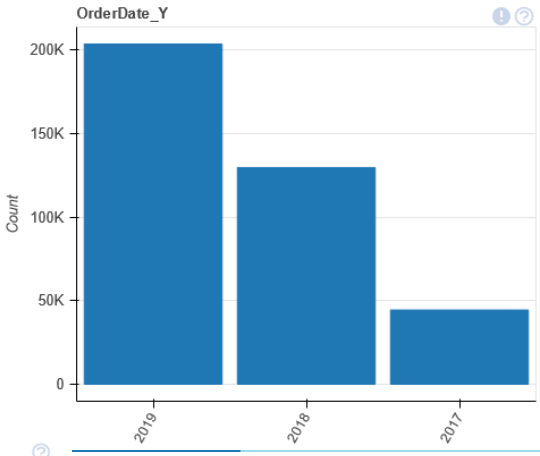


Кроме того, в 30,7% заказов прибыль равна 0. Т.е. почти 43% продаж за весь период не приносили прибыли. Общая прибыль на низком уровне.

Value	Count	Frequency (%)
0.0	116074	30.7%

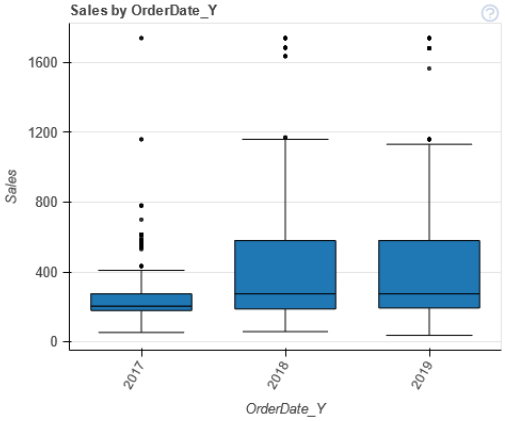
Количество заказов с каждым годом увеличивалось:

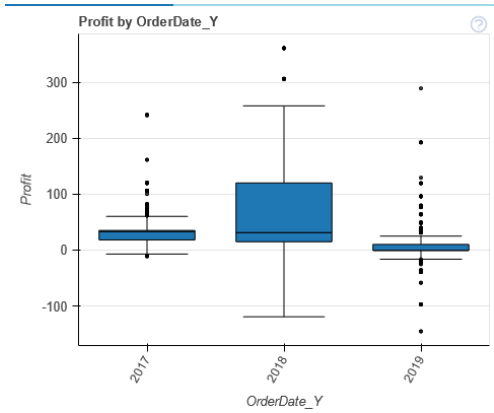
- 2017 - 11,8% всех продаж,
- 2018 - 34,8% всех продаж,
- 2017 - 53,9% всех продаж.



При этом прибыль в 2019 году резко упала, несмотря на рост количества заказ**ов** и суммы продаж.

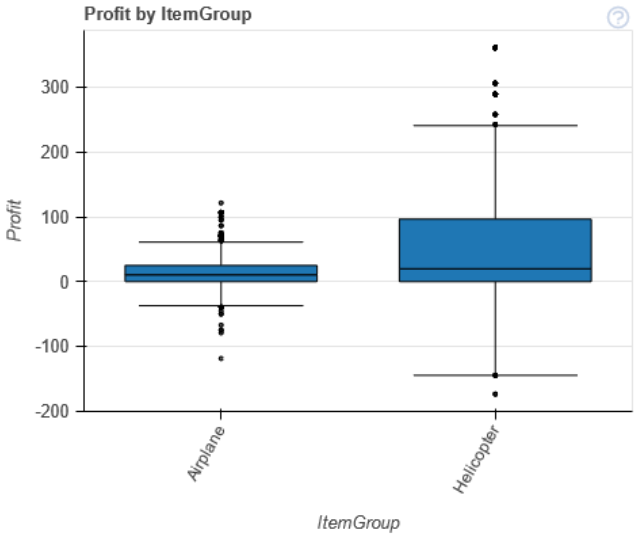
OrderDate_Y	Sales	Profit
2017	10794381.50	1698684.2
2018	50176442.65	7906131.4
2019	80728860.50	2075935.4

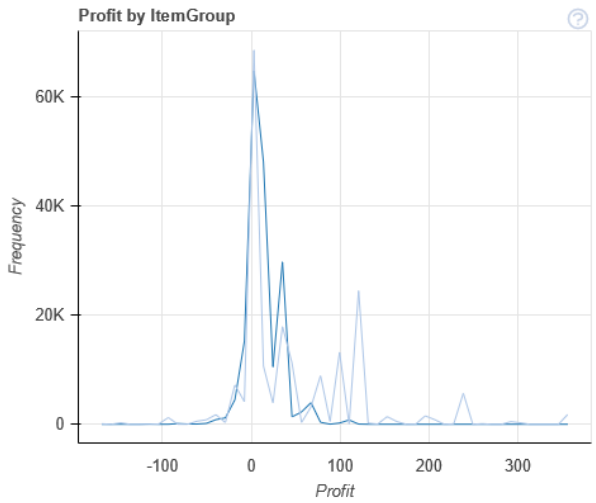




Прибыль была максимальной с апреля 2018 по март 2019 гг. С апреля 2019 года прибыли стала отрицательной при увеличении продаж.

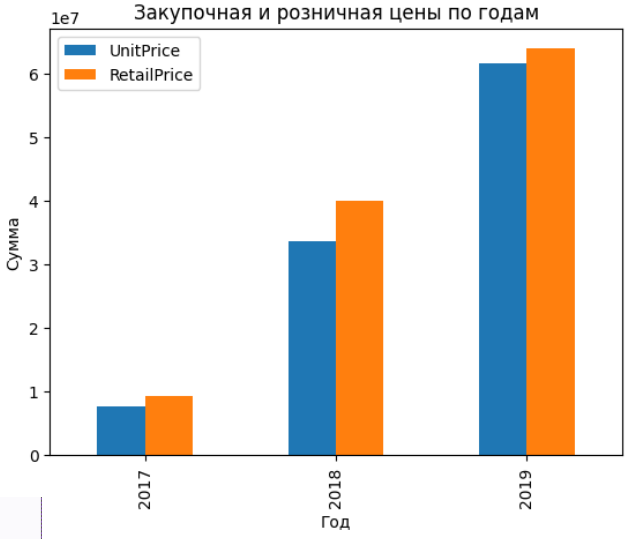
Группа товаров «Вертолеты» более популярны, чем «Самолеты», приносит больше продаж и прибыли. Но в среднем прибыль невысокая, немного выше 0.

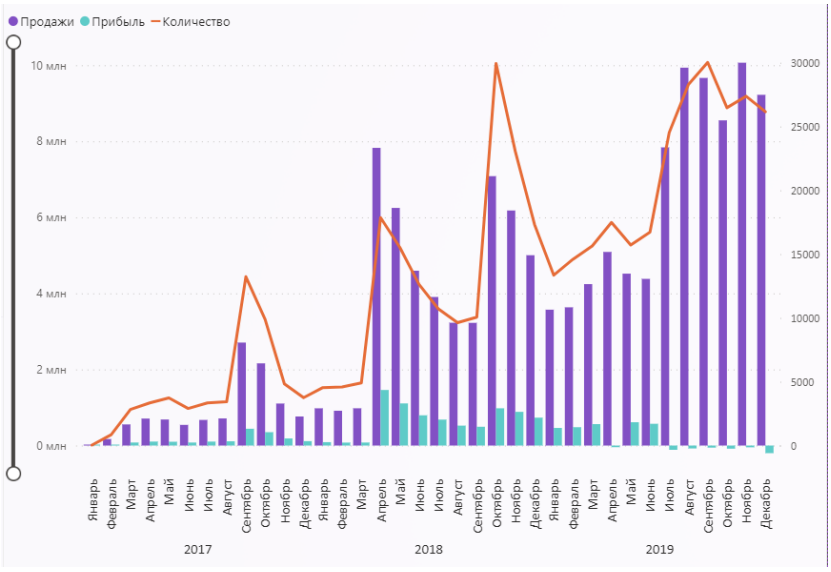




Большая часть прибыли получена за вертолеты в 2018 году.

Розничные цены максимально превышают закупочные в 2018 году, что объясняет величину прибыли в 2018 году.





Также наблюдается сезонность продаж: пик приходится на апрель и сентябрь-октябрь.

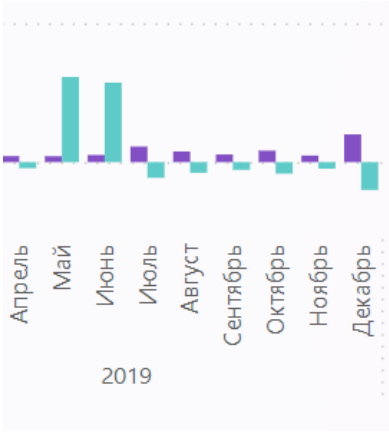
Уже в 2017 году товар с ID 5 продавался в убыток.

В 2018 году уже 3 товара с ID 5, 6, 7 имели отрицательную разницу (созданы закладки в PowerBI).

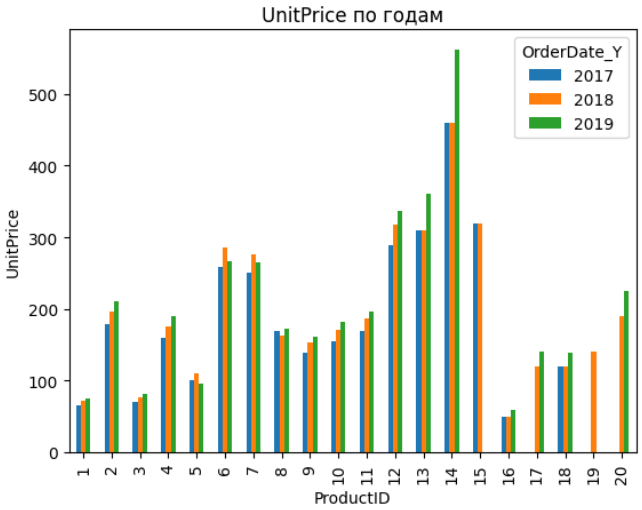
В 2019 году отрицательной разницы между ценой закупки и продажи нет, но разница между ценой закупки и продажи с каждым годом уменьшалась, и в 2019 г. минимальна. В 2019 году закупочная и розничная цена в среднем практически сравнялись, разница крайне мала для получения высокой прибыли.

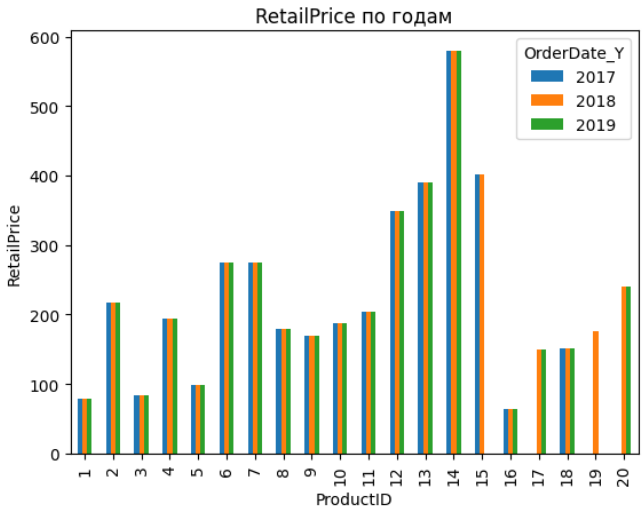
При этом скидки продолжали применяться и в декабре 2019 г. стали максимальными.

Как показал анализ, величина убытка за апрель, июль – декабрь 2019 гг. равна общему размеру скидок соответственно.



Таким образом, закупочная цена менялась, а розничная оставалась на прежнем уровне. Зависимость ценообразования от закупочной цены отсутствует, в связи с чем и провал в прибыли, несмотря на рост продаж.





3. Построение модели машинного обучения

Перед моделированием проведена очистка данных – удалены уникальные, дублирующие, а также высоко коррелированные признаки.

По результатам работы библиотеки AutoGluon лучшей моделью машинного обучения себя показала "WeightedEnsemble_L2" с RMSE = 0.0019.

```
AutoGluon training complete, total runtime = 2117.74s ... Best model: "WeightedEnsemble_L2"
```

На втором месте - "CatBoost": RMSE = 0.0024.

На третьем месте - "XGBoost": RMSE = 0.0077.

На тестовых данных WeightedEnsemble_L2 имеет хорошие показатели:

- RMSE = 0.4075466785870182
- R-squared value = 0.9999991134521942

```
[ ] # Оценка производительности модели
    performance = predictor.evaluate(test_data)

    Evaluation: root_mean_squared_error on test data: -0.4075466785870182
            Note: Scores are always higher_is_better. This metric score can be multiplied by -1 to get the metric value.
    Evaluations on test data:
    {
        "root_mean_squared_error": -0.4075466785870182,
        "mean_squared_error": -0.16609429522731034,
        "mean_absolute_error": -0.0030019108090179316,
        "r2": 0.9999982265731288,
        "pearsonr": 0.9999991134521942,
        "median_absolute_error": -0.00023193359379547474
    }
```

На тестовых данных в тройке лучших моделей остаются те же модели, только в другой последовательности:

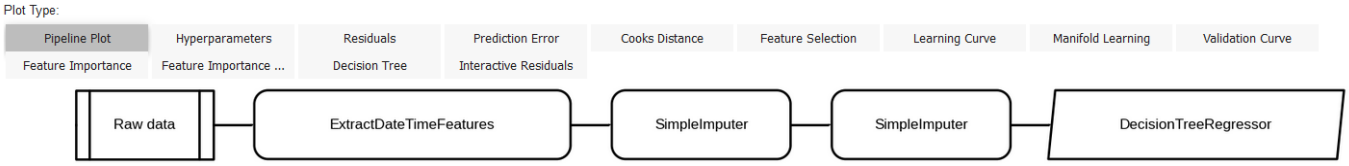
1. CatBoost - RMSE = 0.385243
2. WeightedEnsemble_L2 - RMSE = 0.407547
3. XGBoost - RMSE = 0.531089

```
# Реальные значения целевой переменной в тестовой выборке
y_true = test_data['Sales']

# Подсчет accuracy (R-squared value)
accuracy = (predictions == y_true).mean()
print('Accuracy:', accuracy)

# Сравнение моделей на тестовых данных
leaderboard = predictor.leaderboard(test_data)
```

```
Accuracy: 0.003587075937471045
                  model  score_test   score_val  pred_time_test  pred_time_val
0              CatBoost   -0.385243   -0.002357        0.675379       0.048367
1   WeightedEnsemble_L2   -0.407547   -0.001858      154.489060       5.650414
2               XGBoost   -0.531089   -0.007674       21.572438       1.043764
3         ExtraTreesMSE   -0.714194   -0.029508        1.300524       0.189786
4       RandomForestMSE   -0.789186   -0.010323        1.124733       0.108595
5              LightGBM   -0.803926   -0.113458        1.040536       0.046393
6         LightGBMLarge   -0.985501   -0.032494      132.216944       4.553720
7       NeuralNetFastAI   -2.000201   -1.962827        0.597022       0.055037
8        NeuralNetTorch  -31.972086  -32.539416        0.283699       0.022509
9            LightGBMXT  -46.328757  -47.493372       12.625645       0.630955
10       KNeighborsUnif -324.670120 -319.656965      121.422530       4.285863
11       KNeighborsDist -324.671054 -319.656965      127.735225       3.751364
```

Лучшая модель по результатам работы библиотеки PyCaret - DecisionTreeRegressor.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dt	Decision Tree Regressor	0.0034	0.1665	0.3448	1.0000	0.0010	0.0000	0.2550
et	Extra Trees Regressor	0.0025	0.1105	0.2506	1.0000	0.0006	0.0000	0.4090
rf	Random Forest Regressor	0.0061	0.1915	0.3698	1.0000	0.0010	0.0000	0.7270
lightgbm	Light Gradient Boosting Machine	0.0575	0.3657	0.4926	1.0000	0.0012	0.0002	0.5790
xgboost	Extreme Gradient Boosting	0.0070	0.1111	0.2578	1.0000	0.0005	0.0000	0.4810
gbr	Gradient Boosting Regressor	1.0901	6.1839	2.4799	0.9999	0.0067	0.0038	0.3470
ada	AdaBoost Regressor	40.0431	2258.9833	47.3143	0.9758	0.2788	0.2356	0.2740
knn	K Neighbors Regressor	21.3421	5587.4939	74.7384	0.9400	0.1401	0.0469	26.1770
br	Bayesian Ridge	61.4775	8741.9369	93.4931	0.9062	0.5379	0.2216	0.3400
ridge	Ridge Regression	61.4777	8741.9369	93.4931	0.9062	0.5379	0.2216	0.3490
lr	Linear Regression	61.4785	8741.9369	93.4931	0.9062	0.5379	0.2216	3.5530
lasso	Lasso Regression	61.0865	8780.8686	93.7009	0.9057	0.5628	0.2214	0.3690
llar	Lasso Least Angle Regression	60.5979	8827.7005	93.9501	0.9052	0.5771	0.2187	0.2020
en	Elastic Net	77.3142	19014.6201	137.8816	0.7959	0.2813	0.1835	0.1910
par	Passive Aggressive Regressor	101.6261	20217.2057	139.5477	0.7837	0.5089	0.4007	0.3360
omp	Orthogonal Matching Pursuit	122.6669	37241.3279	192.9698	0.6003	0.3270	0.2700	0.2240
huber	Huber Regressor	79.3180	47195.9652	217.2285	0.4935	0.3586	0.1092	0.2000
dummy	Dummy Regressor	224.8023	93169.8891	305.2235	-0.0001	0.7929	1.0094	0.3300
lar	Least Angle Regression	275.9734	325259.5859	328.2142	-2.3976	0.7677	1.2117	0.2020



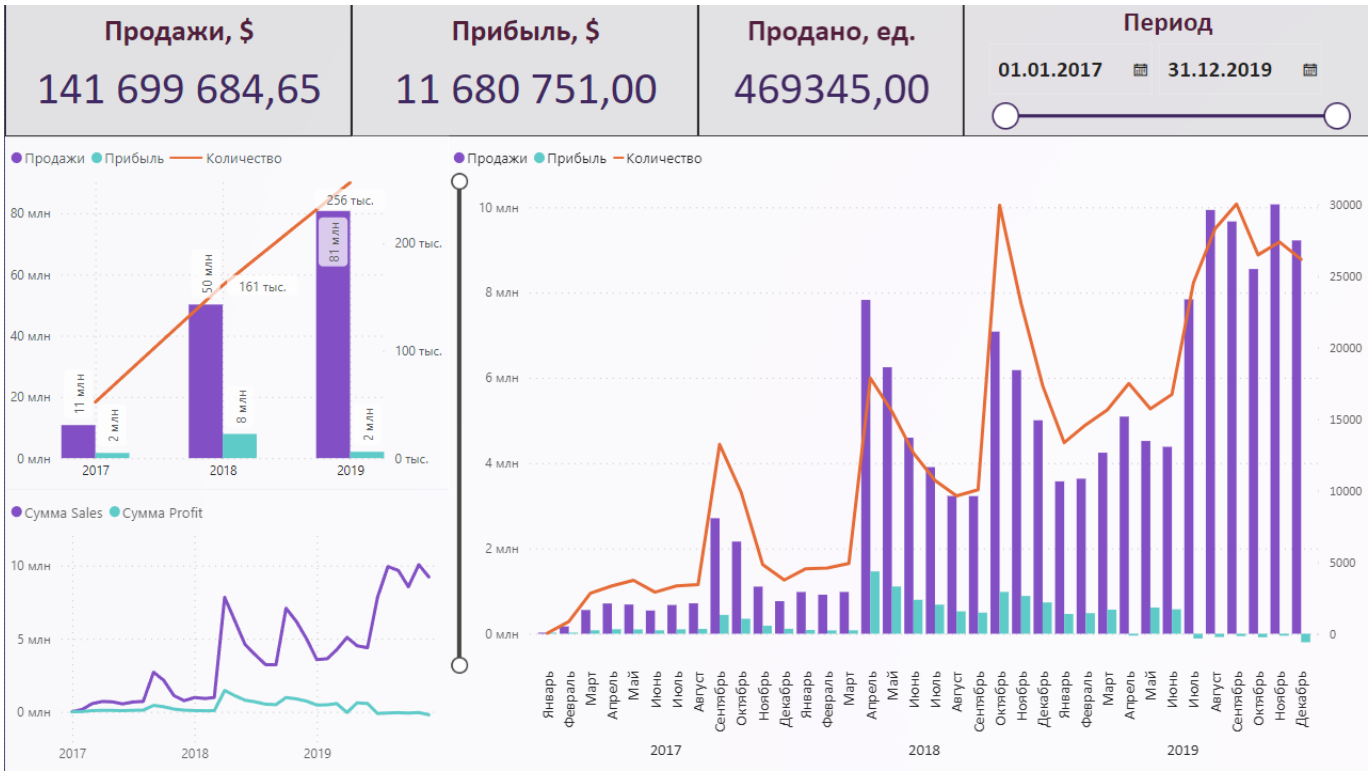

```
# Обучение лучшей модели
trained_model = finalize_model(best_model)

# Оценка на тестовых данных
predictions = predict_model(trained_model, data=test_data)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Decision Tree Regressor	0.0061	1.0545	1.0269	1.0000	0.0016	0.0000

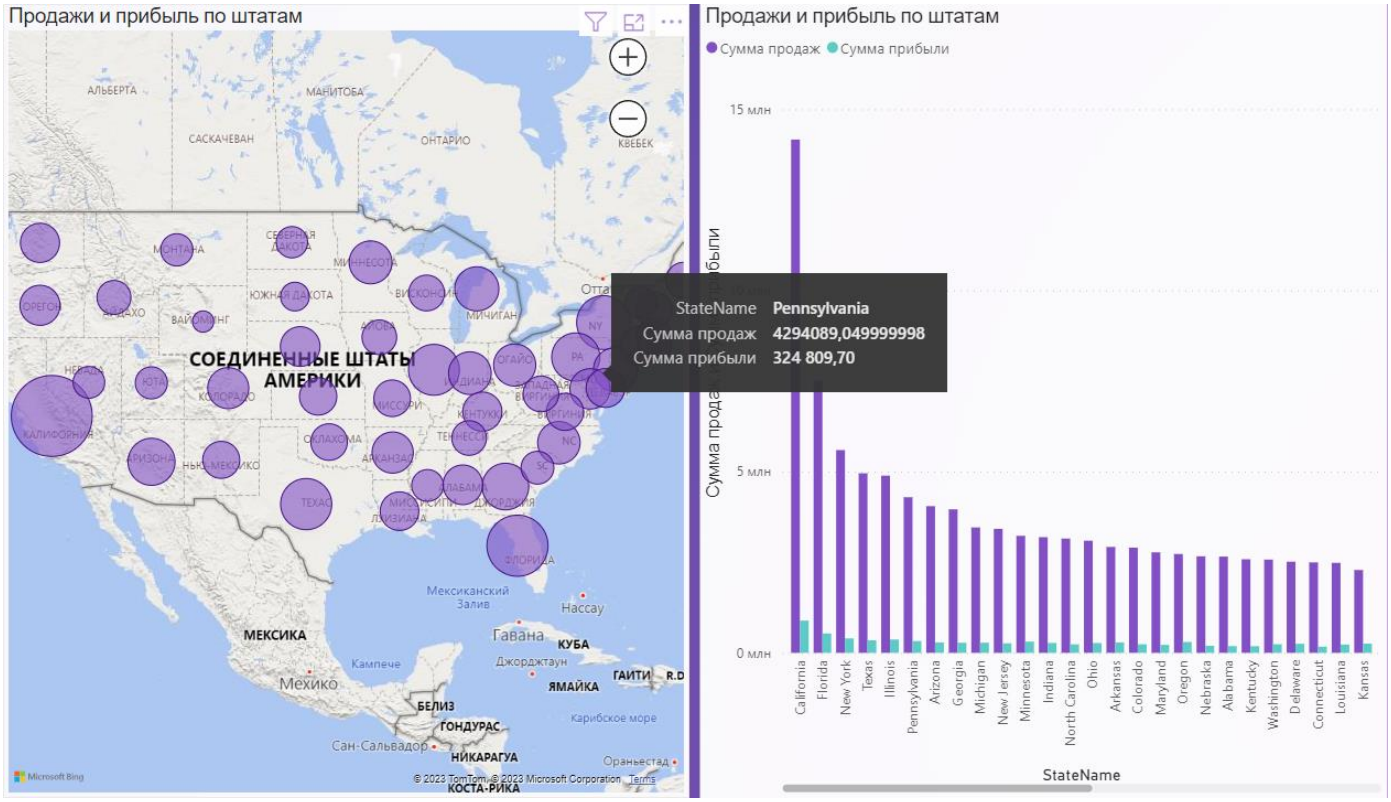
4. Визуализация данных в PowerBI

В PowerBI Desktop разработано несколько отчетов:

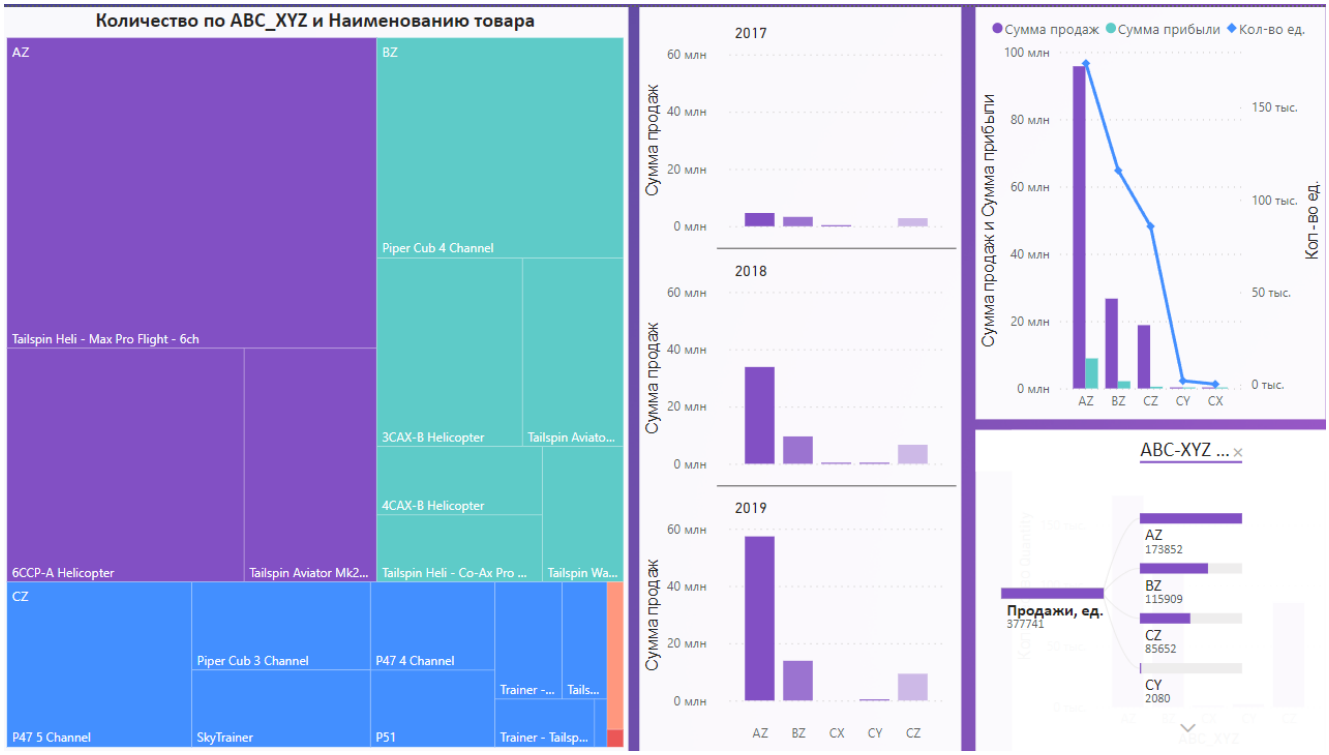
1. Показатели продаж (количество товаров, сумма) и прибыли по годам



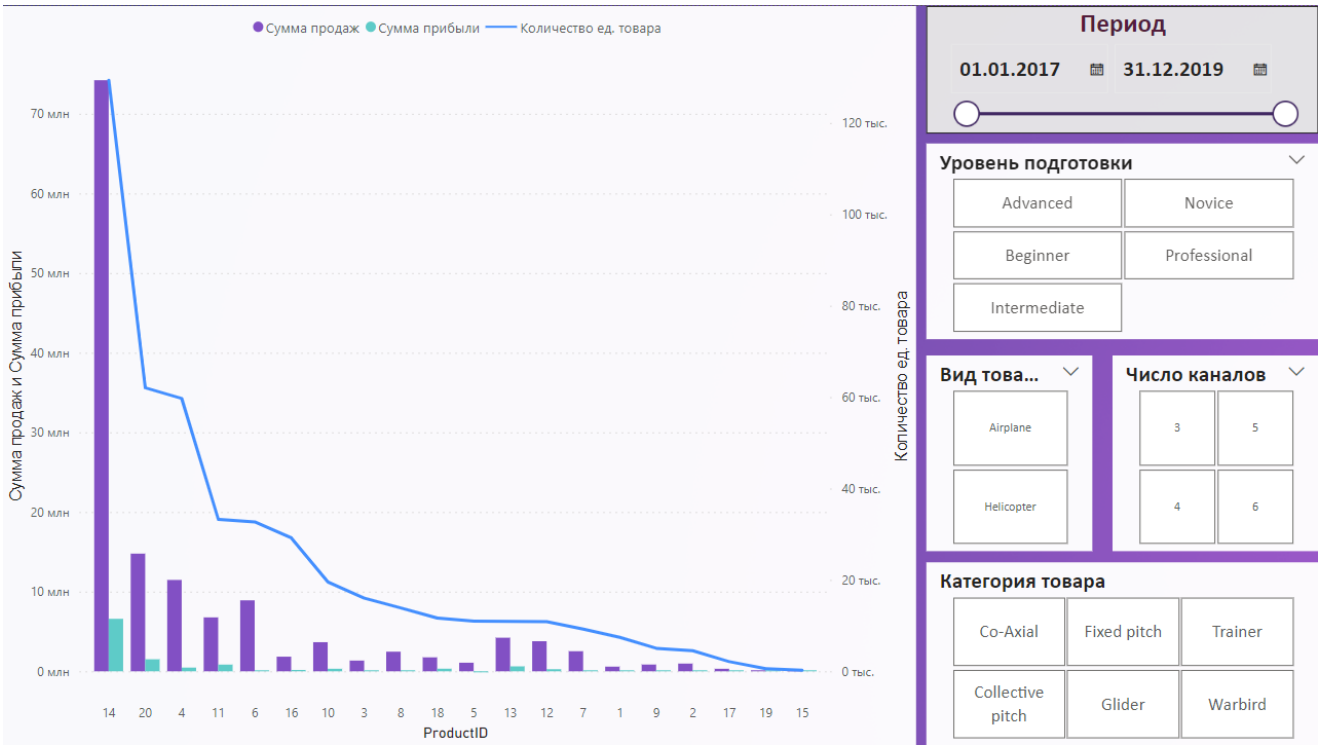
2. Распределение продаж и прибыли по штатам



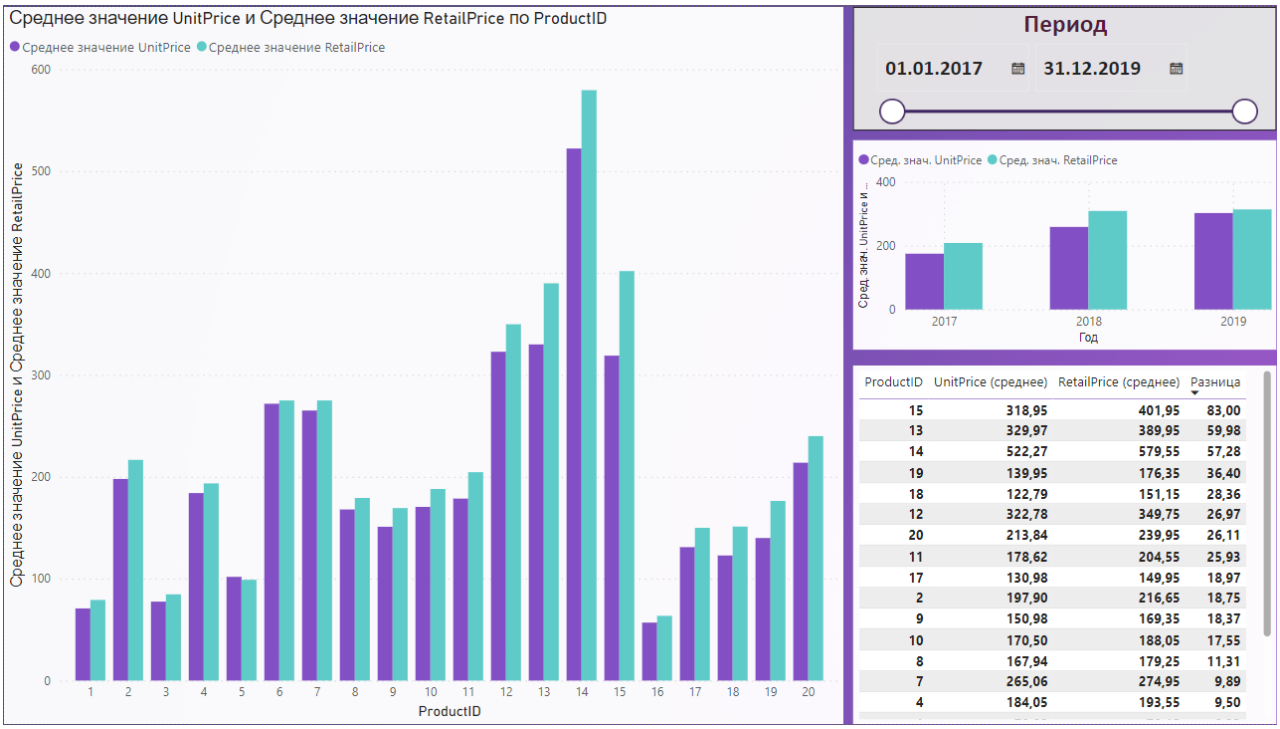
3. Результаты ABC-XYZ анализа: количество проданного товара, сумма продаж и прибыли по каждой категории, а также распределение конкретных товаров по категориям.



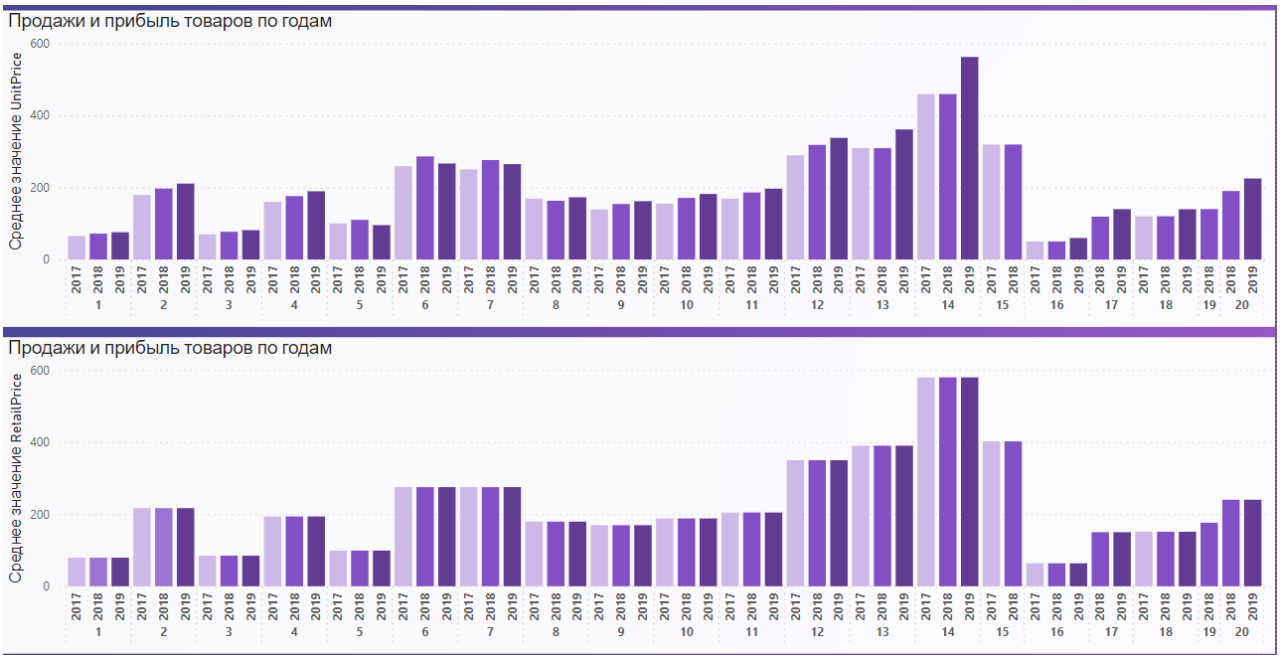
4. Доля продаж (по сумме и количеству) и прибыли по наименованиям (ID), виду, категории, уровню использования товара, количеству каналов и периоду реализации.



5. Сравнение закупочной и розничной цен по годам, ID товаров



ProductID	UnitPrice (среднее)	RetailPrice (среднее)	Разница
15	318,95	401,95	83,00
13	329,97	389,95	59,98
14	522,27	579,55	57,28
19	139,95	176,35	36,40
18	122,79	151,15	28,36
12	322,78	349,75	26,97
20	213,84	239,95	26,11
11	178,62	204,55	25,93
17	130,98	149,95	18,97
2	197,90	216,65	18,75
9	150,98	169,35	18,37
10	170,50	188,05	17,55
8	167,94	179,25	11,31
7	265,06	274,95	9,89
4	184,05	193,55	9,50



6. Соотношение прибыли и скидок

Год	Месяц	UnitPrice	RetailPrice	Discount	Profit
2019	Апрель	4 049 571,55	4 049 571,55	43 359,30	-43 359,30
2019	Май	3 123 551,55	3 643 715,45	42 467,70	613 078,30
2019	Июнь	3 053 429,00	3 549 063,70	51 292,45	572 037,05
2019	Июль	6 178 220,25	6 178 220,25	112 294,90	-112 294,90
2019	Август	7 731 215,40	7 731 215,40	76 829,00	-76 829,00
2019	Сентябрь	7 504 246,65	7 504 246,65	55 246,00	-55 246,00
2019	Октябрь	6 706 034,70	6 706 034,70	82 767,00	-82 767,00
2019	Ноябрь	7 829 491,40	7 829 491,40	47 080,30	-47 080,30
2019	Декабрь	7 325 110,40	7 325 110,40	199 914,00	-199 914,00

Вывод:

1. На размер прибыли (наличие убытков) в данном кейсе повлияло соотношение закупочной и розничной цены.

2. За 3 года закупочная цена менялась в сторону увеличения, а розничная оставалась прежней.

3. Несмотря на отсутствие прибыли продолжали применяться скидки.

4. В апреле 2019 г. и с июля по сентябрь 2019 года розничная цена сравнялась с закупочной, а сумма убытков образовалась в размере скидок.

Решение:

менять ценообразование для получения прибыли.