Национальный исследовательский университет

«Высшая школа экономики»

Московский институт электроники и математики им. А.Н. Тихонова, Москва

Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по

Дисциплине «Методы анализа стохастических взаимосвязей»

Бригада №18:

Антонов Егор Алексеевич, 3 курс, БПМ214

Герасина Дарья Николаевна, 3 курс, БПМ214

Москва 2024

Оглавление

[**1. Общая постановка задачи 3**](#_tamj3p2pxlz5)

[1.1 Описание прикладной области и данных 3](#_xhi88chwx0ct)

[1.2 Основные гипотезы, которые планируется проверить в рамках исследования 4](#_b4xwptrbk0zv)

[**2. Предварительный анализ собранных данных 4**](#_j0ymgyv6b21i)

[2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы 4](#_gfyxkj2gqher)

[2.1.1 Анализ количественных переменных 4](#_j3ierchdqto1)

[2.1.2. Анализ качественных переменных. 6](#_3fn1khngp197)

[2.2 Анализ статистической связи. 9](#_qrsbdz1hhajz)

[2.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная». 9](#_nfeux22wjvqd)

[2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная». 11](#_8dh6x75lzflk)

[2.2.3 Анализ статистической взаимосвязи между независимыми переменными 12](#_6896029w0qi0)

[Качественная независимая переменная - качественная независимая переменная 12](#_2n0o9p3q6m3a)

[Качественная независимая переменная - числовая независимая переменная 14](#_9i1gdjghus1j)

[Числовая независимая переменная - числовая независимая переменная 15](#_thsfhg65swki)

[2.2.4 Предварительная проверка гипотез 16](#_ndw5x1m7p0mq)

[**3. Проверка гипотез с помощью моделирования 16**](#_qrlqza6882bx)

[3.1. Построение базовой модели. 16](#_gh2886lxi4s)

[3.2. Анализ наличия гетероскедастичности 18](#_9a1nfmlg20ay)

[3.3. Проверка мультиколлинеарности 18](#_9oybt8ax800e)

[3.2. Проверка гипотез с помощью моделирования 19](#_2exa52y17q06)

[3.3. Оптимизация итоговой модели, сравнение качества моделей. 20](#_m8v87lb5jrco)

[3.4. Проверка прогностических способностей модели 20](#_tr7wdhzddail)

[**4. Заключение 21**](#_9ooc0yqdih8z)

#### 

#### **1. Общая постановка задачи**

##### 1.1 Описание прикладной области и данных

Для исследования были выбраны данные о стоимости авиабилетов на рейсы 2022 года, вылетающие из аэропортов России.

Данные можно найти по ссылке: <https://www.kaggle.com/datasets/francuzovd/airplane-tickets-from-moscow-2022>

| № | Характеристика объекта/явления | Название переменной | Шкала измерения (одна из  четырех) | Роль:  целевая/  объясняющая |
| --- | --- | --- | --- | --- |
| 1 | Авиакомпания (10 уникальных значений) | airline | номинальная | объясняющая |
| 2 | Количество пересадок (от 0 до 5 шт) | number\_of\_changes | номинальная | объясняющая |
| 3 | Страна прилета (10 уникальных значений) | destination\_  country\_code | номинальная | объясняющая |
| 4 | День недели поиска авиабилета (от 0 до 6) | search\_weekday | номинальная | объясняющая |
| 5 | День недели перелета (от 0 до 6) | depart\_weekday | номинальная | объясняющая |
| 6 | Количество дней между покупкой билета и вылетом (от 0 до 324) | diff\_days | относительная | объясняющая |
| 7 | Расстояние перелета (от 102.71 км до 12588.7 км) | distance | относительная | объясняющая |
| 8 | Цена авиабилета (от 888 руб до 99 895 руб) | price | относительная | целевая |

Таблица 1. Описание факторов, учтенных в анализе.

* airline\_name – 10 уникальных значений, некоторые из них: “Turkish Airlines”, “Wizz Air”, “S7 Airlines”, “Aeroflot”, “Red Wings Airlines”
* destination\_country\_code – 10 уникальных значений, некоторые из них: “RU” – Россия, “TR” – Турция, “IT” – Италия, “US” – Соединенные Штаты

##### 1.2 Основные гипотезы, которые планируется проверить в рамках исследования

Простая гипотеза: при увеличении расстояния(distance), увеличивается и значение цены авиабилета (целевая переменная price)

Сложные гипотезы:

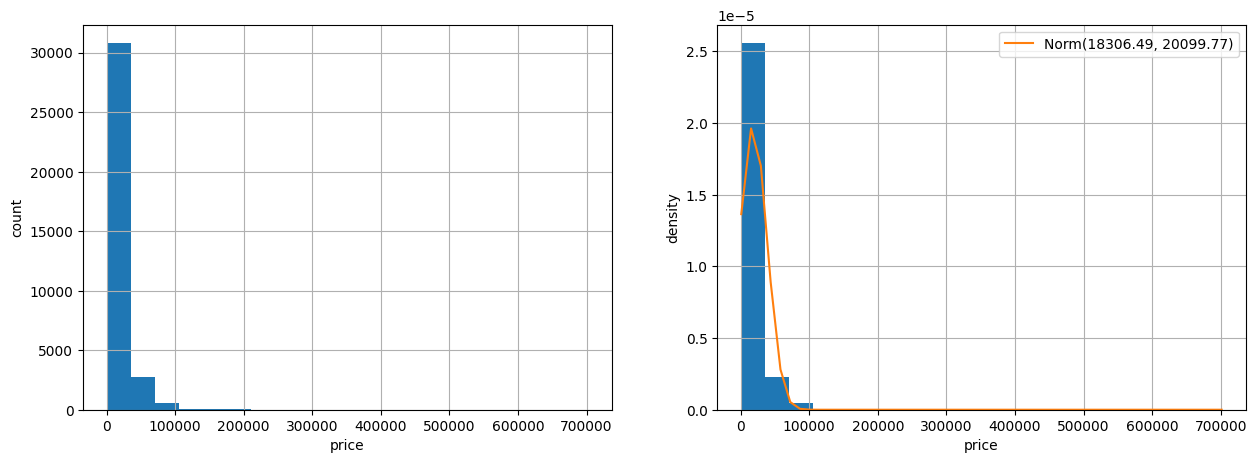
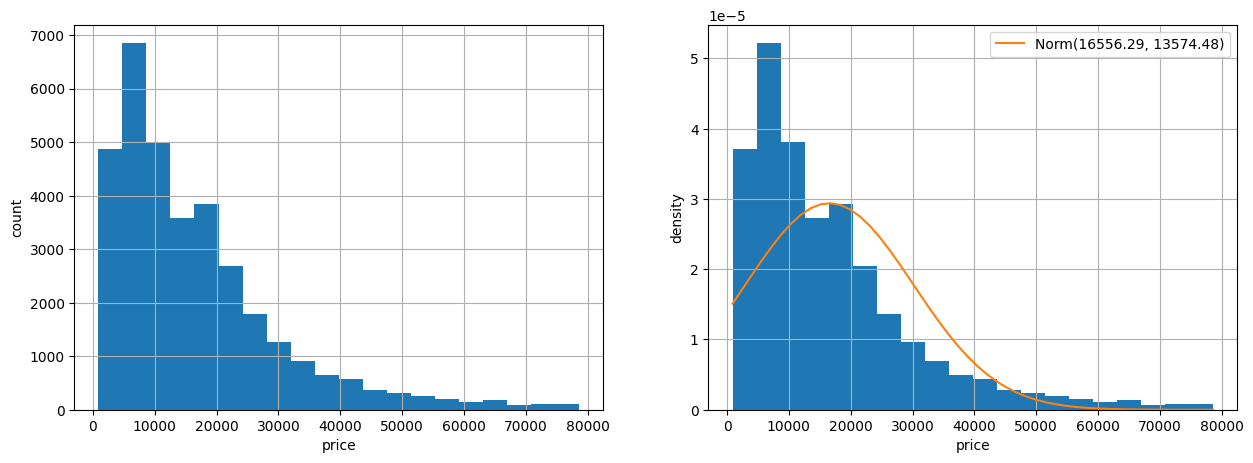
1. В течение определенного количества дней с момента покупки до вылета стоимость авиабилетов увеличивается, после чего остается на том же уровне.
2. С увеличением расстояния, скорость роста цены авиабилета на внутренних рейсах ниже чем на зарубежных.

#### **2. Предварительный анализ собранных данных**

##### 2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

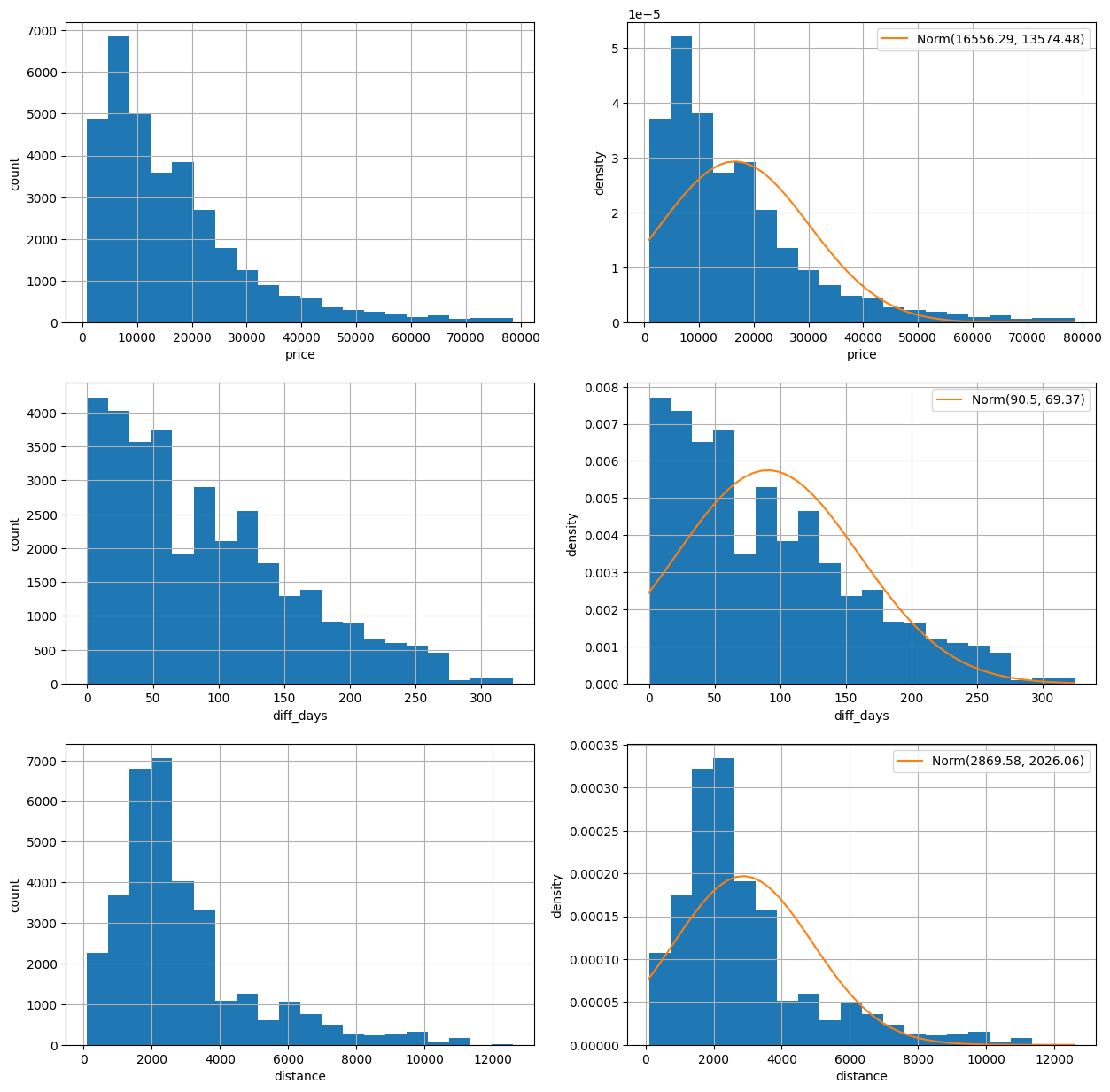
###### 2.1.1 Анализ количественных переменных

**Стоимость авиабилета (price)**

*рисунок 1. Гистограммы целевой переменной price до удаления выбросов и после*

| Статистика | Значение Price | Значение diff\_days | Значение distance |
| --- | --- | --- | --- |
| Среднее | 16556.28 | 90.4 | 2869.57 |
| Медиана | 12699 | 75 | 2263.35 |
| Стандартное отклонение | 13574.48 | 69.37 | 2026.06 |
| Межквартильный размах | 15116 | 98 | 1816.37 |
| Верхняя квартиль | 21830 | 132 | 3420.17 |
| Нижняя квартиль | 6714 | 34 | 1603.80 |
| Коэффициент асимметрии | 1.695 | 0.814 | 1.627 |
| Коэффициент эксцесса | 3.34 | -0.073 | 2.795 |
| Минимальное значение | 888 | 0 | 102.71 |
| Максимальное значение | 78575 | 324 | 12588.7 |
| Количество наблюдений | 33795 | 33795 | 33795 |
| Количество пропущенных значений | 0 | 0 | 0 |

*Таблица 2. Статистики числовых переменных*



*рисунок 2. Гистограммы переменных price, distance, diff\_days*

На гистограммах видно наличие выбросов, избавимся от них, с помощью метода трех сигм

На основании анализа гистограмм, представленных выше, можно сделать вывод о том, что распределение стоимости ассиметрично вправо. Это подтверждается также таблицей описательных статистик: медиана (12699) < среднего (16556.28), коэффициент асимметрии (1.695) положительный.

Это можно объяснить тем, что цена билета на авиарейс не может быть отрицательной, а значит, она ограничена минимальным значением в 0. Однако, максимальной цены на билет может не быть, и она может быть довольно высокой, особенно для бизнес-класса или в периоды повышенного спроса.

Также на рынке авиабилетов часто есть ограниченное количество билетов по самой низкой цене. После того как эти билеты будут проданы, цены начинают расти, что приводит к дальнейшей асимметрии вправо в распределении цен

Коэффициент эксцесса положительный (3.34) => можем говорить об островершинности распределения

**Переменная ‘Количество дней между покупкой билета и вылетом’ (diff\_days)**

На основании анализа гистограмм, представленных выше, можно сделать вывод о том, что распределение количества дней между покупкой билета и вылетом ассиметрично вправо. Это подтверждается также таблицей описательных статистик: медиана (75) < среднего (90.4), коэффициент асимметрии (0.814) положительный.

Это можно объяснить тем, что кол-во дней до вылета не может быть отрицательной, а значит, она ограничена минимальным значением в 0. Однако, максимальное значения может не быть, и оно может быть довольно высокой, так как некоторые люди любят покупать билеты сильно заранее (так как часто это бывает довольно выгодно).

Коэффициент эксцесса отрицательный (-0.073) => у распределения нет острых пиков

**Переменная distance**

На основании анализа гистограмм, представленных выше, можно сделать вывод о том, что распределение расстояния перелета ассиметрично вправо. Это подтверждается также таблицей описательных статистик: медиана (2263.35) < среднего (2869.57), коэффициент асимметрии (1.627) положительный.

Это объясняется тем, что большинство авиарейсов, особенно на популярных маршрутах, составляют короткие или средние расстояния. Такие рейсы востребованы для деловых поездок и туризма, что приводит к концентрации значений на нижнем уровне расстояний.

Также, дальние международные рейсы, как правило, менее часты по сравнению с короткими внутренними рейсами. Они обслуживают более узкий круг пассажиров и, как правило, их меньше, что создает длинный хвост справа в распределении расстояний.

Коэффициент эксцесса (2.795) говорит о наличии острого пика у распределения

###### 2.1.2. Анализ качественных переменных.

**Переменная *search\_weekday* Переменная *depart\_weekday***

| День недели | Частота | % | % допустимых | Накопленный % | День недели | Частота | Проценты | Процент допустимых | Накопленный процент |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| пн | 6635 | 19.6 | 19.6 | 19.6 | пн | 5205 | 15.4 | 15.4 | 15.4 |
| вт | 2857 | 8.5 | 8.5 | 28.1 | вт | 5056 | 15.0 | 15.0 | 30.4 |
| ср | 5415 | 16.0 | 16.0 | 44.1 | ср | 5177 | 15.3 | 15.3 | 45.7 |
| чт | 5426 | 16.1 | 16.1 | 60.2 | чт | 5301 | 15.7 | 15.7 | 61.4 |
| пт | 3937 | 11.6 | 11.6 | 71.8 | пт | 4500 | 13.3 | 13.3 | 74.7 |
| сб | 6736 | 19.9 | 19.9 | 91.7 | сб | 4947 | 14.6 | 14.6 | 89.3 |
| вс | 2789 | 8.3 | 8.3 | 100.0 | вс | 3609 | 10.7 | 10.7 | 100.0 |
| Всего | 33795 | 100.0 | 100.0 | 100.0 | Всего | 33795 | 100.0 | 100.0 | 100 |

*Таблица 2. Статистики качественных переменных search\_weekday и depart\_weekday*

**Переменная *number\_of\_changes* (количество пересадок)**

| Количество пересадок | Частота | % | % допустимых | Накопленный % | Количество пересадок | Частота | % | % допустимых | Накопленный % |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 8298 | 24.6 | 24.6 | 24.6 | 0 | 8298 | 24.6 | 24.6 | 24.6 |
| 1 | 15378 | 45.5 | 45.5 | 70.1 | 1 | 15378 | 45.5 | 45.5 | 70.1 |
| 2 | 7854 | 23.2 | 23.2 | 93.3 | 2 | 7854 | 23.2 | 23.2 | 93.3 |
| 3 | 2131 | 6.3 | 6.3 | 99.6 | 3+ | 2256 | 6.7 | 6.7 | 100.0 |
| 4 | 132 | 0.4 | 0.4 | 100.0 |  |  |  |  |  |
| 5 | 2 | 0.0 | 0.0 | 100.0 |  |  |  |  |  |
| Всего | 33795 | 100.0 | 100.0 | 100.0 | Всего | 33795 | 100.0 | 100.0 | 100.0 |

*Таблица 3. Статистики качественной переменной number\_of\_changes до укрупнения и после*

Менее чем у 5% всех перелетов пересадок больше чем 3, потому закономерно будет объединить категории “3”, “4” и “5” в одну категорию “3+”.

**Переменная *airline* (авиакомпания)**

| Категория | Частота | % | % допустимых | Накопленный % | Категория | Частота | % | % допустимых | Накопленный % |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S7 Airlines | 5565 | 16.5 | 16.5 | 16.5 | S7 Airlines | 5565 | 16.5 | 16.5 | 16.5 |
| Azimut | 4984 | 14.74 | 14.74 | 31.24 | Azimut | 4984 | 14.74 | 14.74 | 31.24 |
| Aeroflot | 4407 | 13.04 | 13.04 | 44.28 | Aeroflot | 4407 | 13.04 | 13.04 | 44.28 |
| Turkish Airlines | 3976 | 11.76 | 11.76 | 56.04 | Turkish Airlines | 3976 | 11.76 | 11.76 | 56.04 |
| Pobeda | 3836 | 11.35 | 11.35 | 67.39 | Pobeda | 3836 | 11.35 | 11.35 | 67.39 |
| Uzbekistan Airways | 3521 | 10.41 | 10.41 | 77.8 | Uzbekistan Airways | 3521 | 10.41 | 10.41 | 77.8 |
| Nordwind Airlines | 2555 | 7.56 | 7.56 | 85.36 | Nordwind Airlines | 2555 | 7.56 | 7.56 | 85.36 |
| Pegasus Airlines | 2482 | 7.34 | 7.34 | 92.7 | Pegasus Airlines | 2482 | 7.34 | 7.34 | 92.7 |
| Wizz Air | 1337 | 3.95 | 3.95 | 96.65 | Other | 2469 | 7.3 | 7.3 | 100 |
| KLM | 1132 | 3.35 | 3.35 | 100 |  |  |  |  |  |
| Всего | 33795 | 100 | 100 |  | Всего | 33795 | 100 | 100 | 100 |

*Таблица 4. Статистики качественной переменной airline до укрупнения и после*

Две авиакомпании “Wizz Air” и “KLM” набрали менее 5%. Объединим их в категорию “Others”

**Переменная destination\_country\_code (страна прилета)**

| Категория | Частота | % | % допустимых | Накопленный % | Категория | Частота | % | % допустимых | Накопленный % |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RU | 17397 | 51.48 | 51.48 | 51.48 | RU | 17397 | 51.48 | 51.48 | 51.48 |
| TR | 3331 | 9.85 | 9.85 | 61.33 | Tourism | 7257 | 21.45 | 21.45 | 72.93 |
| UZ | 2563 | 7.58 | 7.58 | 68.91 | Culture | 5139 | 15.19 | 15.19 | 88.12 |
| DE | 1980 | 5.85 | 5.85 | 74.76 | CIS | 4002 | 11.83 | 11.83 | 100 |
| IT | 1854 | 5.48 | 5.48 | 80.24 |  |  |  |  |  |
| KZ | 1439 | 4.25 | 4.25 | 84.49 |  |  |  |  |  |
| IN | 1426 | 4.21 | 4.21 | 88.7 |  |  |  |  |  |
| ES | 1366 | 4.04 | 4.04 | 92.74 |  |  |  |  |  |
| US | 1305 | 3.86 | 3.86 | 96.6 |  |  |  |  |  |
| GR | 1134 | 3.35 | 3.35 | 100 |  |  |  |  |  |
| Всего | 33795 | 100 | 100 | 100 | Всего | 33795 | 100 | 100 | 100 |

*Таблица 5. Статистики качественной переменной destination\_country\_code до укрупнения и после*

Многие категории содержат менее 5% значений. Можно разделить страны, основываясь на географических, культурных и экономических признаках:

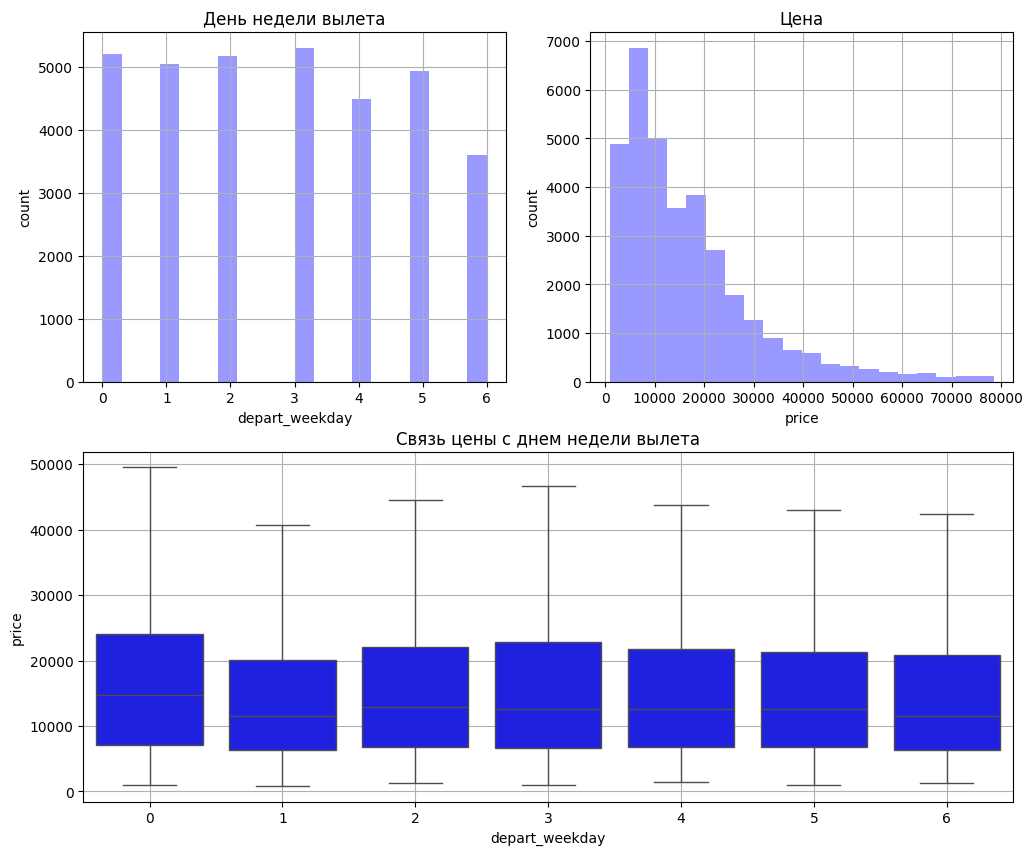
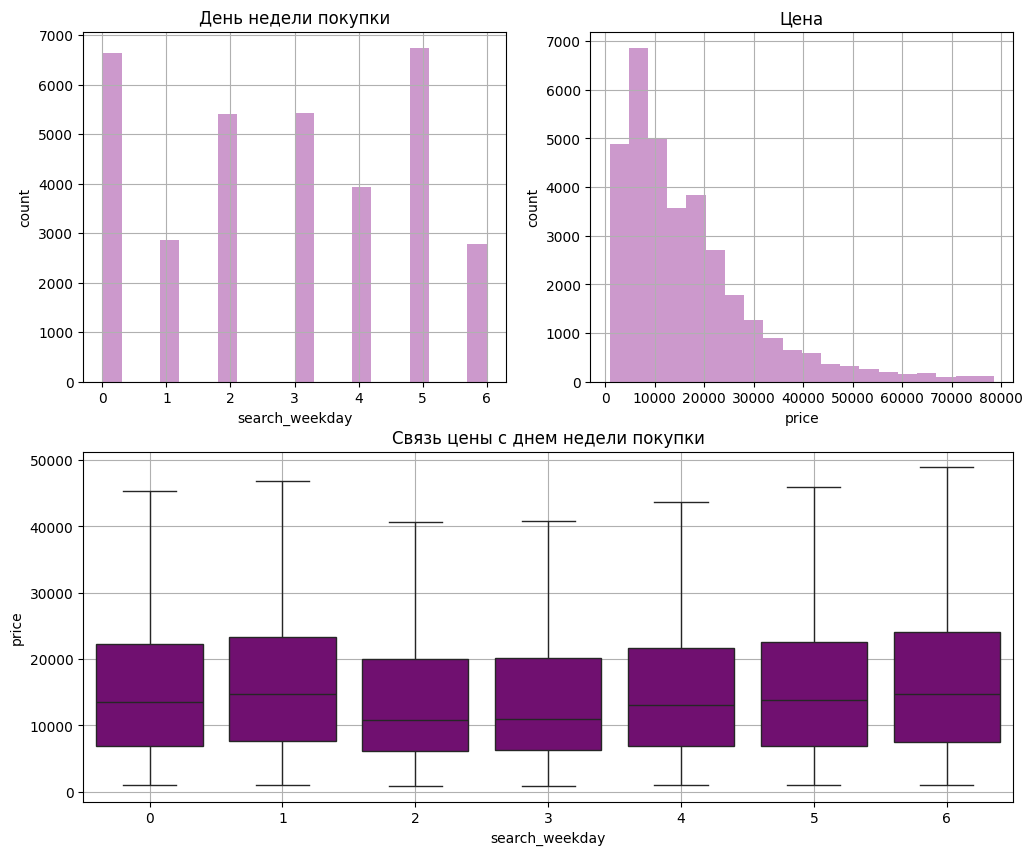
1. Domestic – RU (Россия) – оставляем отдельно, т. к. это внутренние перелеты
2. Tourism – TR (Турция), ES (Испания), GR (Греция), IN (Индия) – популярные туристические направления
3. Culture – DE (Германия), IT (Италия), US (США) – страны с сильными экономическими и культурными связями
4. CIS – UZ (Узбекистан), KZ (Казахстан) – бывшие советские республики и страны с историческими связями

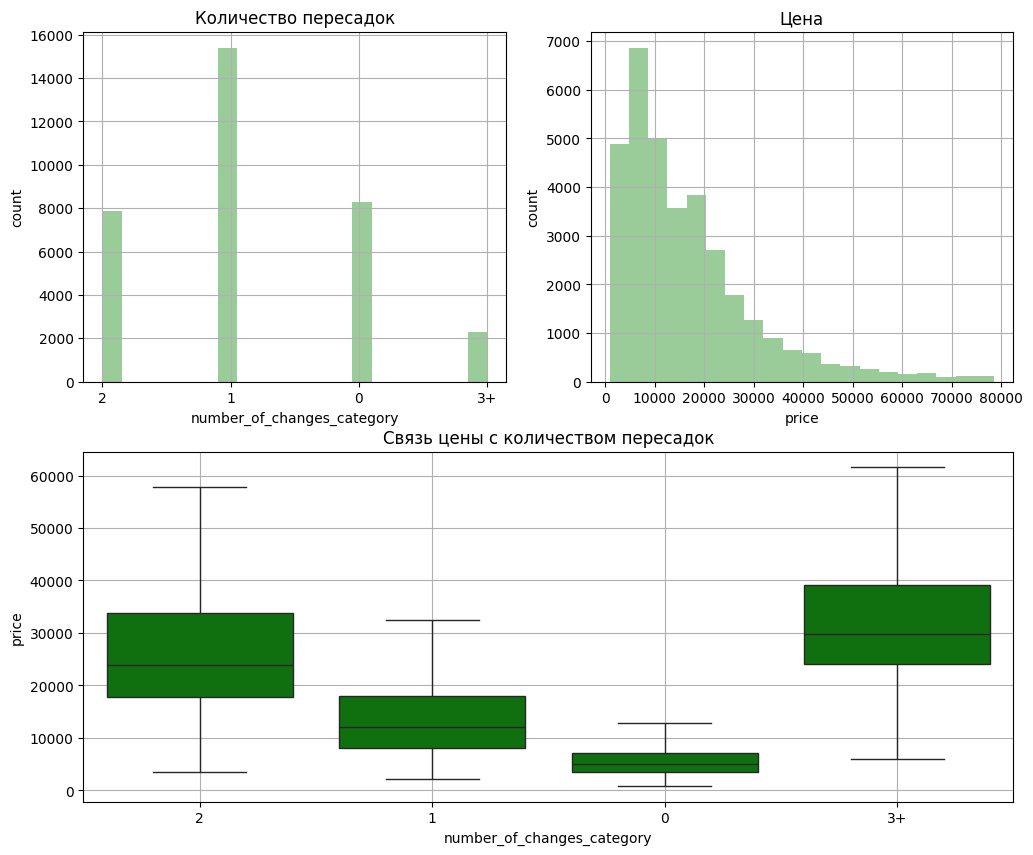
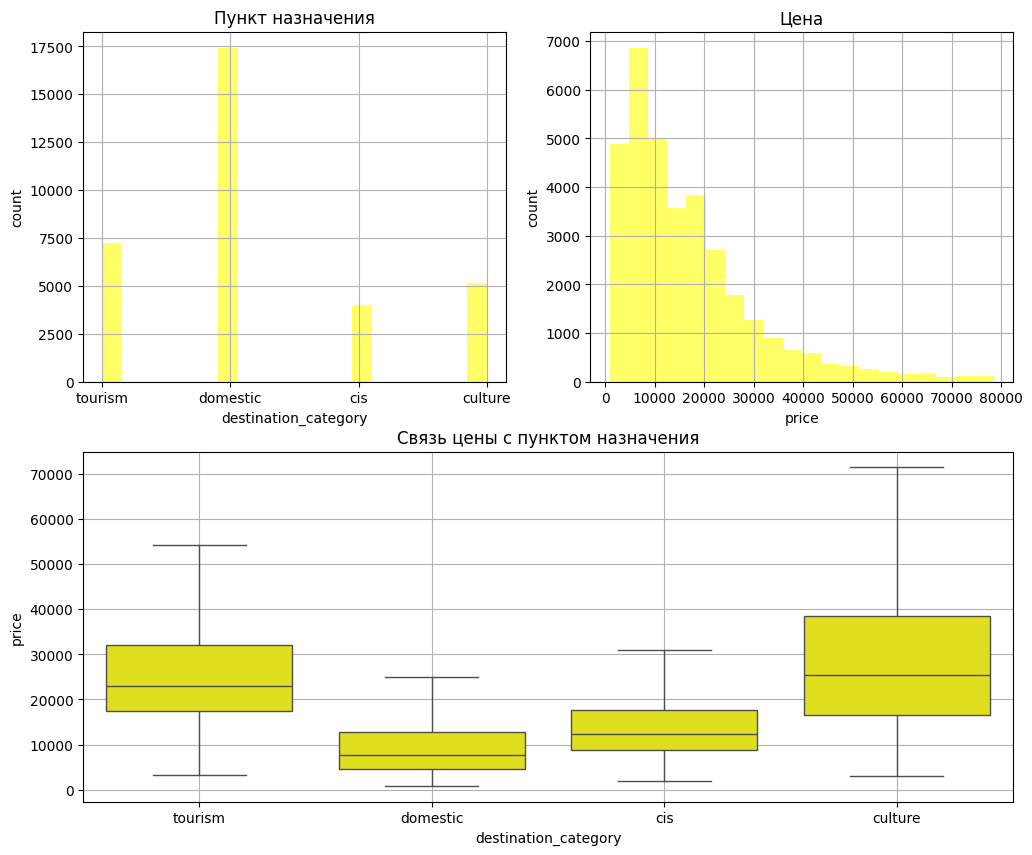
##### 

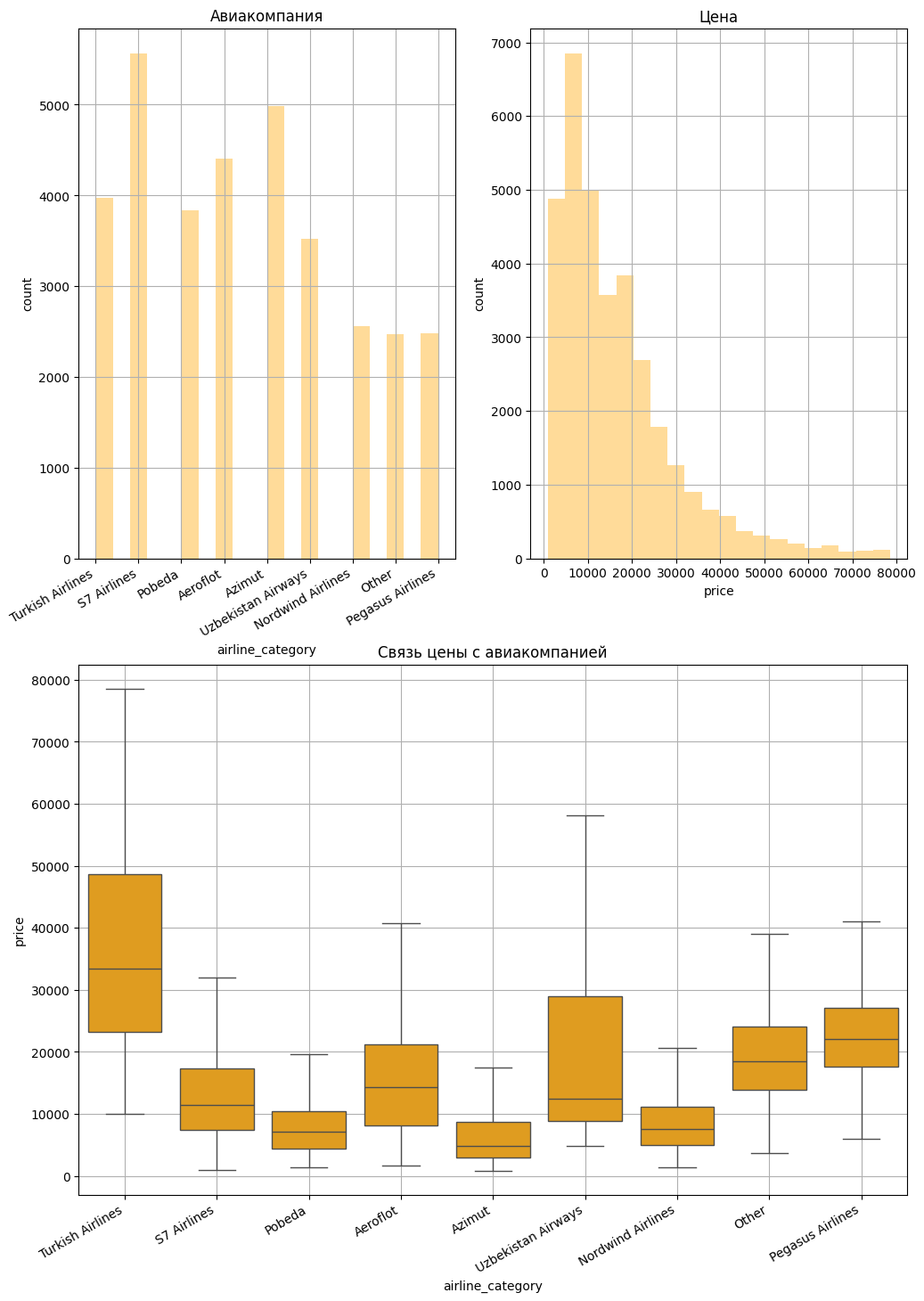
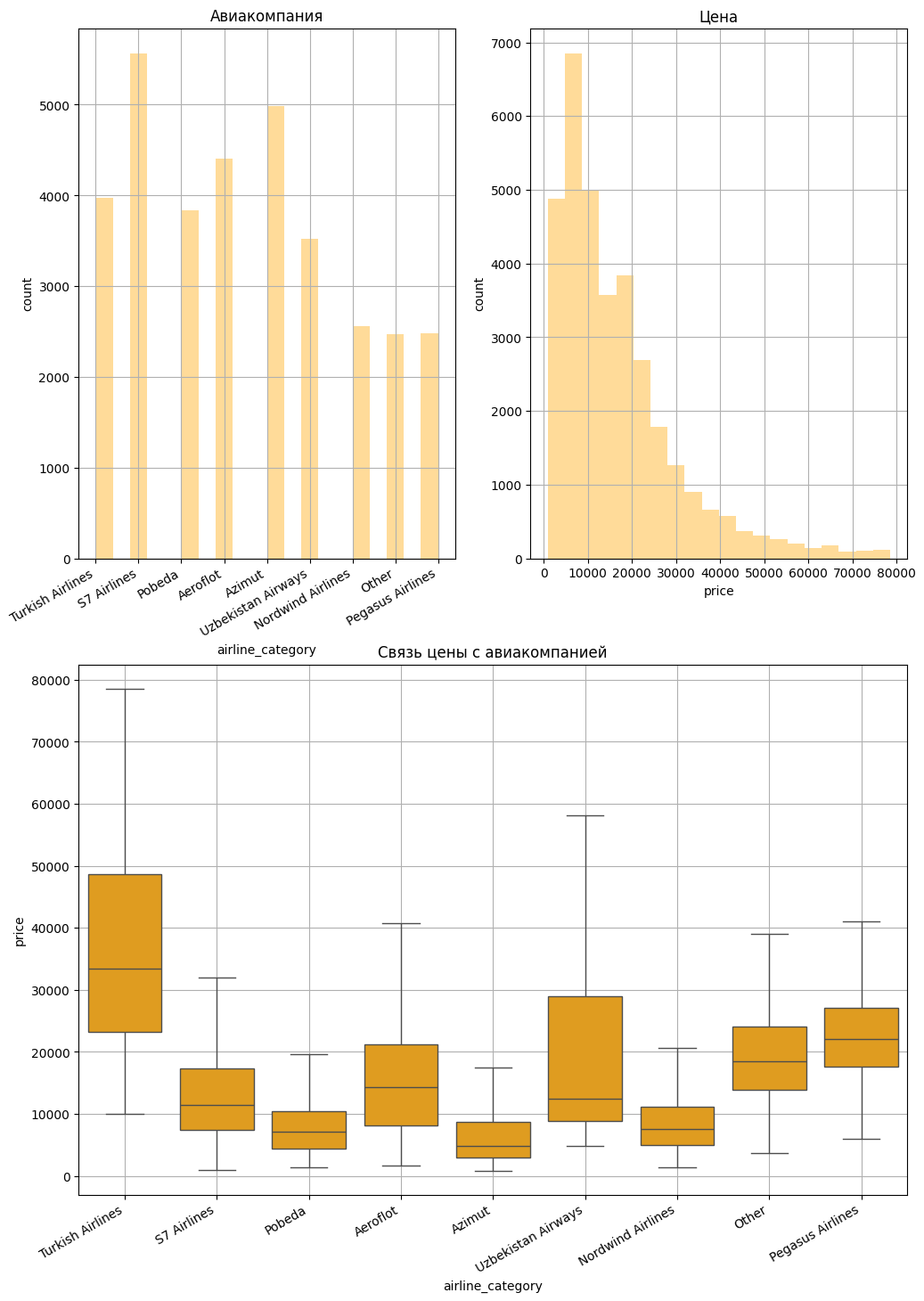
*рисунок 2. Столбчатые диаграммы качественных переменных*

##### **2.2 Анализ статистической связи.**

###### 2.2.1 Графический анализ пары «целевая переменная – качественная объясняющая переменная».





*рисунок 3. диаграммы связи категориальных переменных с целевой*

**Зависимость цены авиабилета (price) от дня недели покупки авиабилета (search\_weekday)**

Проанализировав диаграмму пары переменных “стоимость авиабилета” – “день недели покупки билета”, можно сделать вывод, что цены могут изменяться в зависимости от дня недели покупки. Наличие данной взаимосвязи подтверждается тестом Крускала-Уоллиса: p-value = 5.58e-57, что меньше определенного порога 0.05, следовательно, нулевая гипотеза об отсутствии взаимосвязи отвергается. Данный эффект может объясняться тем, что в разные дни недели покупатели могут действовать по-разному: например, в будние дни бизнесмены покупают больше билетов, а на выходных — туристы, что может влиять на ценообразование.

**Зависимость цены (price) от дня недели вылета (depart\_weekday)**

Анализируя диаграмму пары переменных стоимость авиабилета – день недели вылета, можно сделать вывод, что цена авиабилета зависит от дня недели вылета. Наличие данной взаимосвязи подтверждается тестом Крускала–Уоллиса: p-value = 6.41e-27 < 0.05, а следовательно, отвергается нулевая гипотеза об отсутствии взаимосвязи. Цены могут быть выше в выходные и праздничные дни из-за повышенного спроса со стороны путешественников.

**Зависимость цены (price) от количества пересадок (number\_of\_changes\_category)**

Из анализа диаграммы Бокса-Уискера для пары переменных стоимость – количество пересадок, можно сделать вывод, что цена авиабилета варьируется в зависимости от количества пересадок. Используя критерий Крускала-Уоллиса, мы получили значение p-value = 0.0 < 0.05, что говорит нам о необходимости отвержения нулевой гипотезы об отсутствии влияния качественной переменной на целевую. Билеты с пересадками, как правило, дороже прямых рейсов, так как дальность перелета зачастую больше.

**Зависимость цены (price) от пункта назначения (destination\_category)**

На основе анализа диаграммы пары переменных стоимость – пункт назначения, можно сделать вывод о взаимосвязи диапазона стоимости и категории пункта назначения, а именно: цены на билеты варьируются в зависимости от пункта назначения. Данная взаимосвязь подтверждается критерием Крускала-Уоллиса: значение p-value = 0.0 < 0.05 – отвергается нулевая гипотеза об отсутствии взаимосвязи. Эта взаимосвязь может быть обоснована тем фактом, что некоторые направления более популярны или находятся дальше, что увеличивает стоимость перелета.

**Зависимость цены (price) от авиакомпании (airline\_category)**

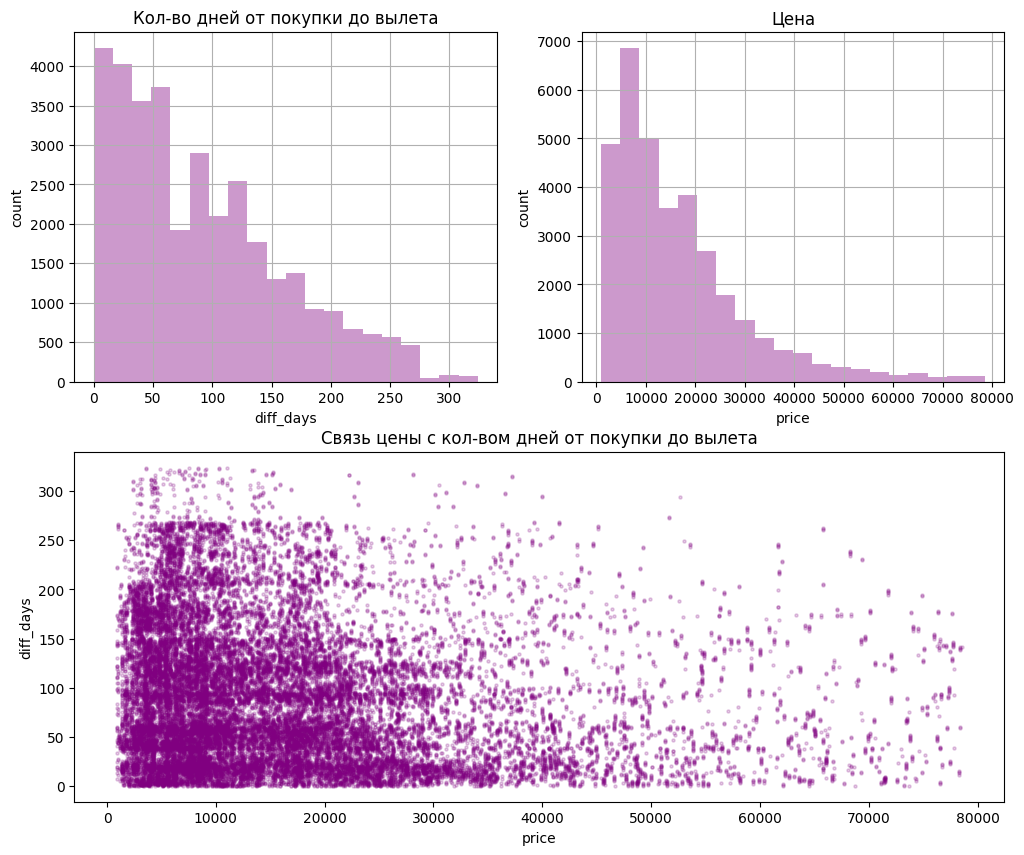
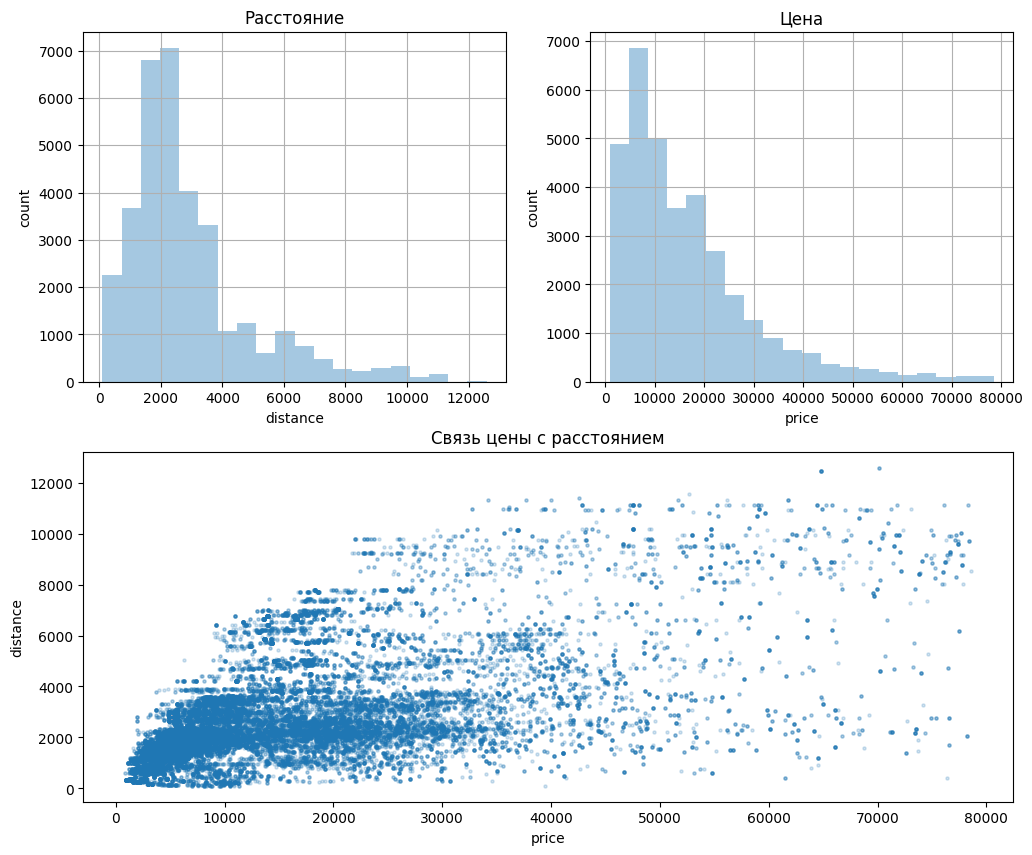
Проанализировав диаграмму пары переменных: стоимость авиабилета – авиакомпания, можно сделать вывод о наличии статистической связи между авиакомпанией и ценой билета. Эта взаимосвязь подтверждается критерием Крускала-Уоллиса: значение p-value = 0 < 0.005 – отвергается нулевая гипотеза об отсутствии взаимосвязи. Разные авиакомпании имеют разную политику ценообразования, что отражается на стоимости билетов.

###### 2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

**Зависимость цены (price) от distance и от diff\_days**

| **price** | Пирсон | Спирмен | Кендалл тау | **diff\_days** | Пирсон | Спирмен | Кендалл тау |
| --- | --- | --- | --- | --- | --- | --- | --- |
| к-т корр. | 0.569 | 0.58 | 0.417 | к-т корр. | -0.157 | -0.172 | -0.116 |
| p-value | 0 | 0 | 0 | p-value | 9.55e-187 | 1.70e-224 | 1.33e-225 |

*Таблица 6. коэффициенты корреляции между целевой и числовыми переменными*

*рисунок 3. диаграммы связи числовых переменных с целевой*

Из диаграммы рассеивания можно сделать предположение о наличии положительной статистической взаимосвязи между целевой переменной *price* и переменной *distance:*  с ростом расстояния цена авиабилета увеличивается.

Возьмем за нулевую гипотезу предположение о том, что между переменными нет никакой статистической взаимосвязи, рассчитаем коэффициенты корреляции и p-value для них. Возьмем уровень статистической значимости = 0.05

Для каждого из методов получили p-value = 0 < 0.05 => нулевая гипотеза о том, что между переменными нет статистической взаимосвязи, отвергается в пользу альтернативной – между price и *distance* существует положительная корреляция

**Зависимость цены (price) от кол-ва дней от покупки билета до вылета (diff\_days)**

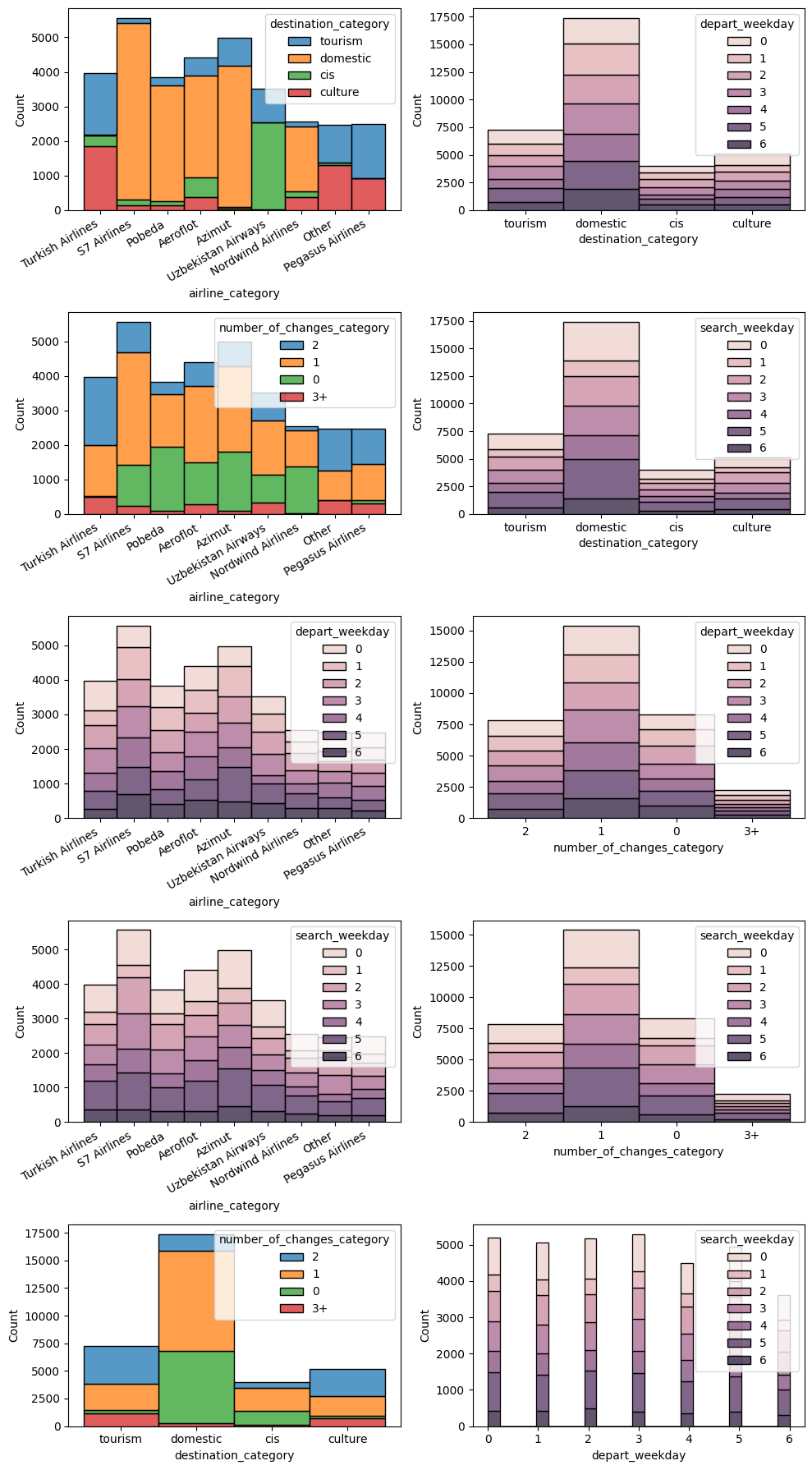
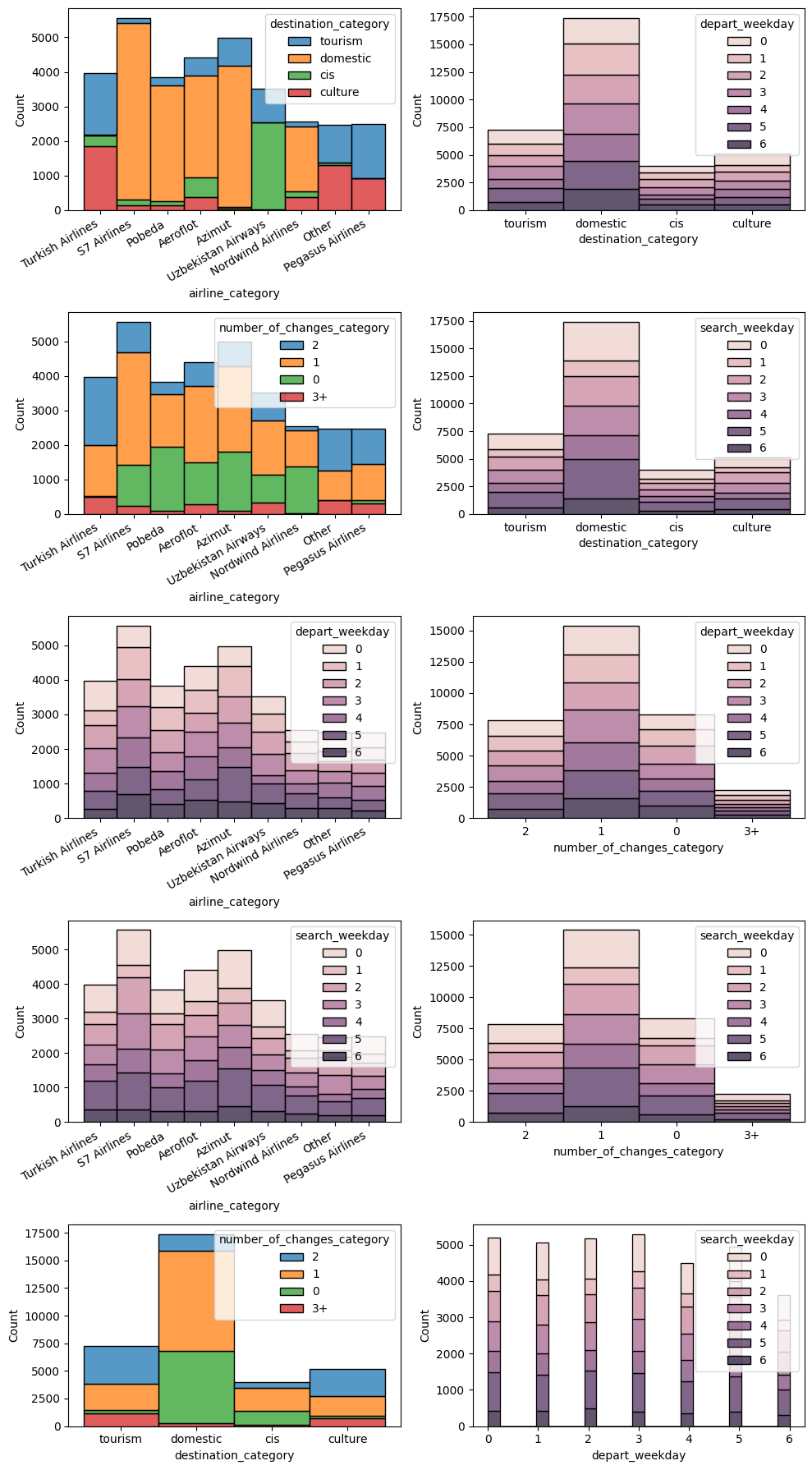
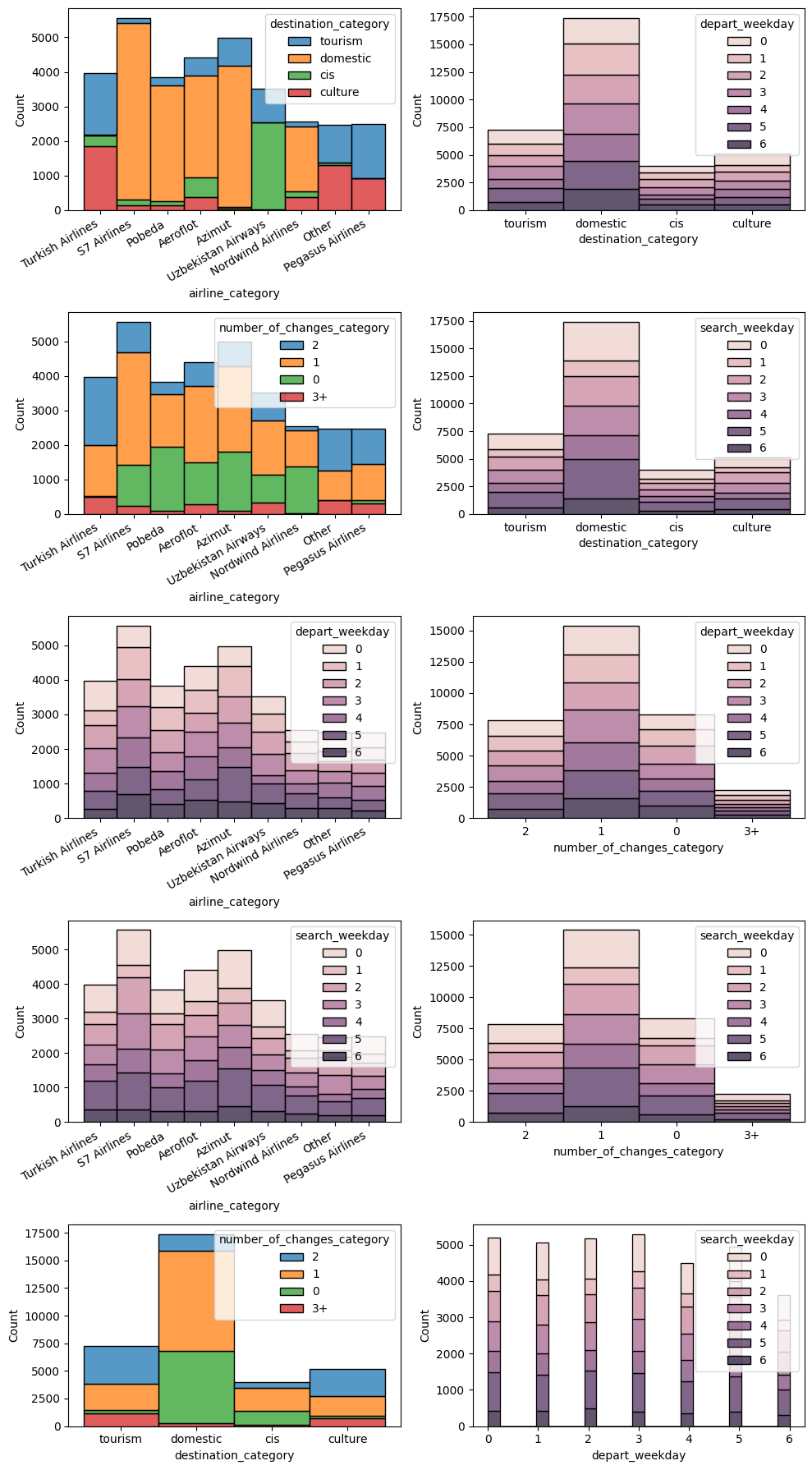
Из диаграммы рассеивания можно сделать предположение о наличии небольшой отрицательной статистической взаимосвязи между целевой переменной *price* и переменной *diff\_days:*  с уменьшением количества дней цена авиабилета увеличивается.

Возьмем за нулевую гипотезу предположение о том, что между переменными нет никакой статистической взаимосвязи, рассчитаем коэффициенты корреляции и p-value для них. Возьмем уровень статистической значимости .

Для каждого из методов получили p-valueнулевая гипотеза о том, что между переменными нет статистической взаимосвязи, отвергается в пользу альтернативной – между price и *diff\_days* существует небольшая отрицательная корреляция

###### 2.2.3 Анализ статистической взаимосвязи между независимыми переменными

###### Качественная независимая переменная - качественная независимая переменная



*рисунок 4. диаграммы связи независимых качественных переменных*

|  | *airline\_category to destination\_category* | *airline\_category to number\_of\_changes\_category* | *destination\_category to number\_of\_changes\_category* |
| --- | --- | --- | --- |
| *Статистика хи-квадрат* | 36319.6603 | 8885.5763 | 11094.5885 |
| *хи-квадрат test p-value* | 0 | 0 | 0 |
| *V Крамера* | 0.5183 | 0.2564 | 0.2865 |

*Таблица 7a. значения статистик для связи качественных переменных*

|  | *airline\_category to depart\_weekday* | *airline\_category to search\_weekday* | *depart\_weekday to search\_weekday* | *destination\_category to depart\_weekday* |
| --- | --- | --- | --- | --- |
| *Статистика хи-квадрат* | 1056.6897 | 535.0889 | 54.2172 | 477.5017 |
| *хи-квадрат test p-value* | 0 | 0 | 0.0262 | 0 |
| *V Крамера* | 0.0668 | 0.0476 | 0.0151 | 0.0594 |

*Таблица 7b. значения статистик для связи качественных переменных*

|  | *destination\_category to search\_weekday* | *number\_of\_changes\_category to depart\_weekday* | *number\_of\_changes\_category to search\_weekday* |
| --- | --- | --- | --- |
| *Статистика хи-квадрат* | 85.3190 | 193.8664 | 176.3867 |
| *хи-квадрат test p-value* | 0 | 0 | 0 |
| *V Крамера* | 0.0251 | 0.0379 | 0.0361 |

*Таблица 7c. значения статистик для связи качественных переменных*

Из анализа диаграмм и статистических тестов выявлена статистически значимая взаимосвязь (p-value близко к 0, V Крамера не слишком мал) в трех случаях:

**авиакомпания - страна назначения** (airline\_category to destination\_category):

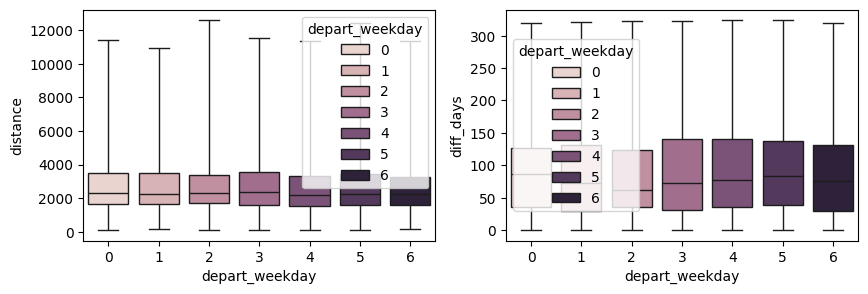
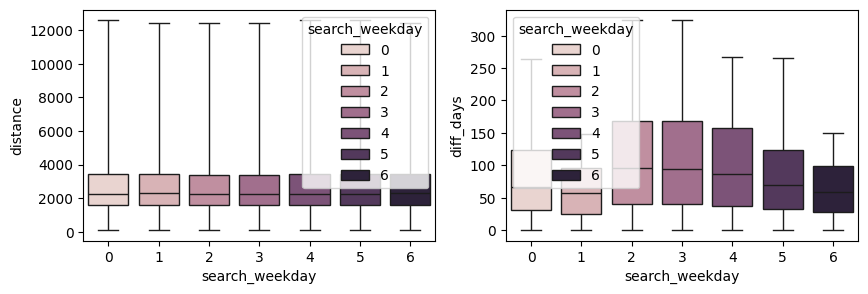
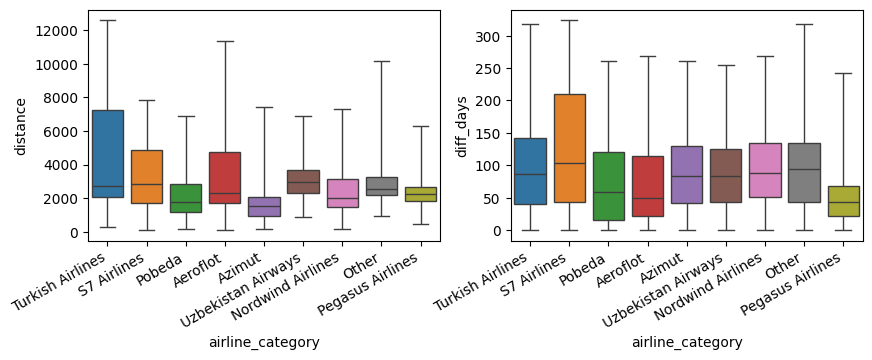
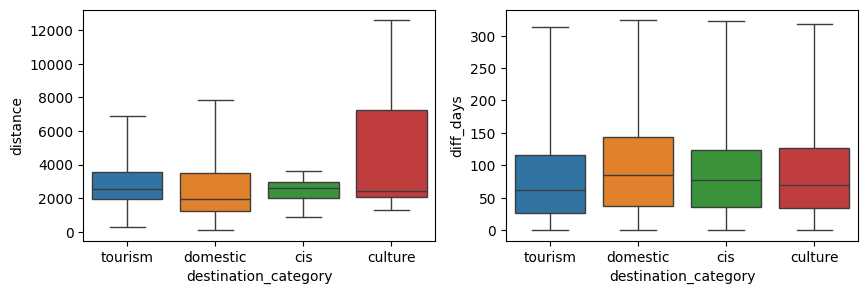
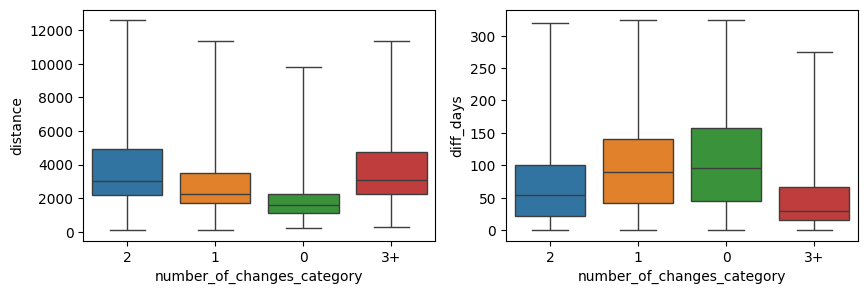
Сила связисредняя (V Крамера = 0.5183). Это объясняется тем, что различные авиакомпании обслуживают разные направления, некоторые авиакомпании могут специализироваться на определенных регионах или маршрутах, что отражается в средней силе связи

**авиакомпания - кол-во пересадок** ( airline\_category to number\_of\_changes\_category): Связь статистически значима, но она слабее чем в предыдущем случае (V Крамера = 0.2564). Объясняется тем, что количество пересадок зависит от политики и логистики авиакомпаний. Некоторые авиакомпании предлагают больше прямых рейсов, а другие — больше рейсов с пересадками.

**страна назначения - кол-во пересадок** (destination\_category to number\_of\_changes\_category): Сила связи примерно такая же как и в предыдущем случае (V Крамера = 0.2865). Происходит из-за того, что направление рейса влияет на количество пересадок. Ближайшие и популярные направления могут иметь больше прямых рейсов, в то время как дальние и менее популярные направления часто требуют пересадок.

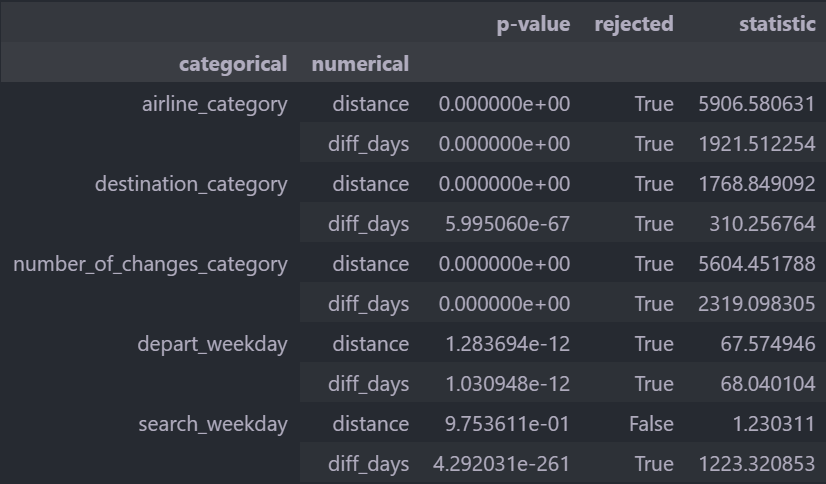
Во всех **остальных** случаях, несмотря на p-value, близкие к нулю, коэффициенты V Крамера очень маленькие, что означает, что зависимости практически нет.

###### Качественная независимая переменная - числовая независимая переменная



###### 

*рисунок 5. диаграммы связи независимых качественных переменных с числовыми*



*Таблица 8. значения статистик для связи качественных переменных с числовыми*

Почти все качественные переменные связаны как с расстоянием перелета, так и с количеством дней между покупкой билета и вылетом (за исключением связи search\_weekday – distance, там мы не смогли отвергнуть нулевую гипотезу об отсутствии связи).

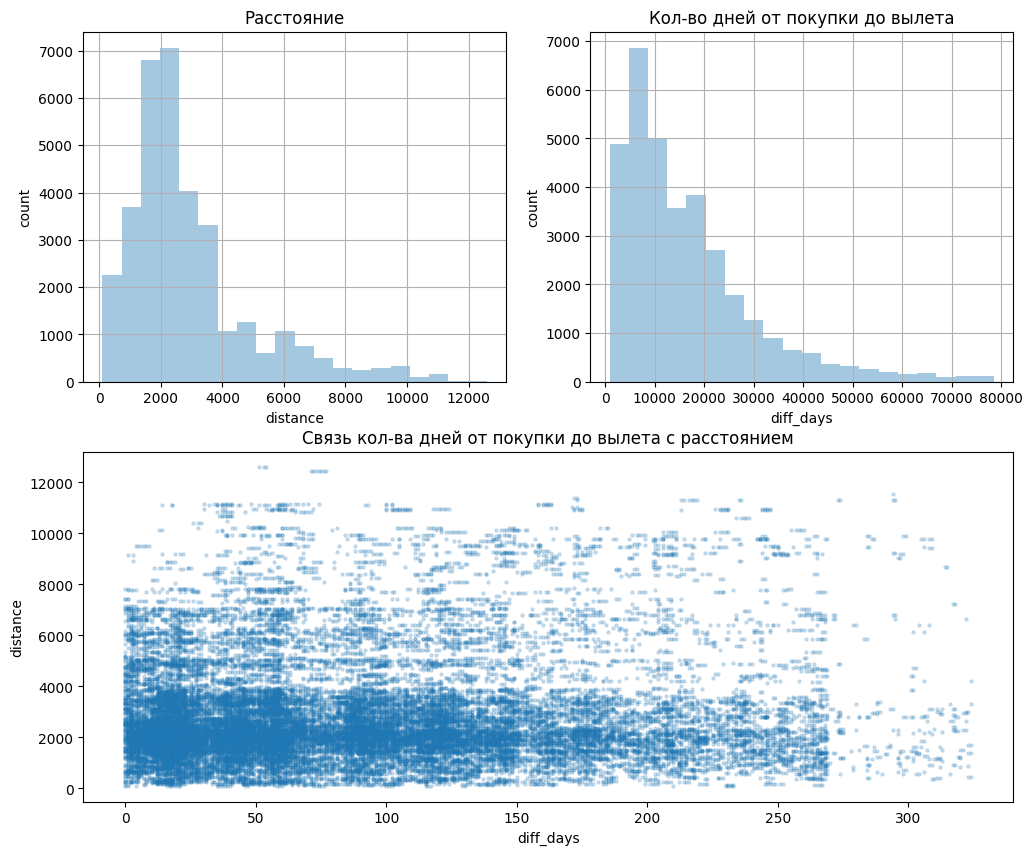
Это подтверждается статистическими тестами, показывающими значимые p-value, что позволяет отвергнуть нулевую гипотезу об отсутствии связи.

Например, *(number\_of\_changes\_category – distance)* больше пересадок обычно связано с более длинными маршрутами, так как прямые рейсы часто короче; *(destination\_category – date\_diff)* международные рейсы обычно бронируются заранее, по сравнению с внутренними рейсами, из-за сложностей планирования и более высоких затрат.

###### Числовая независимая переменная - числовая независимая переменная

|  | Пирсон | Спирмен | Кендалл тау |
| --- | --- | --- | --- |
| Коэффициент корреляции | 0.039 | -0.004 | -0.003 |
| p-value | 0 | 0.39 | 0.38 |
| Нулевая гипотеза | отвергается | не отвергается | не отвергается |

*Таблица 9. значения статистик для связи числовых переменных*

Значение коэффициента Пирсона (0.039) указывает на очень слабую линейную связь между расстоянием перелета (distance) и количеством дней между покупкой билета и вылетом (date\_diff). Однако, p-value равное 0 указывает на статистическую значимость этой связи, что позволяет отвергнуть нулевую гипотезу об отсутствии линейной связи. Возможно, это связано с тем, что люди иногда планируют дальние поездки заранее, но связь слишком слаба, чтобы делать уверенные выводы.

*рисунок 6. связь нез. числовых переменных*

Коэффициенты Спирмена и Кендалла тау показывают практически нулевую корреляцию (-0.004 и -0.003 соответственно), что указывает на отсутствие монотонной зависимости между переменными. Значения p-value (0.39 и 0.38) также не позволяют отвергнуть нулевую гипотезу.

###### 2.2.4 Предварительная проверка гипотез

**Простая гипотеза:** Увеличение расстояния перелета положительно влияет на цену авиабилета.

**Обоснование:** Фактические данные подтверждают, что коэффициенты корреляции между ценой авиабилета и расстоянием перелета (Пирсон = 0.569, Спирмен = 0.58, Кендалл тау = 0.417) все являются положительными и статистически значимыми (p-value = 0 для всех тестов).

Это означает, что чем больше расстояние перелета, тем выше цена авиабилета. Такой вывод поддерживается графическим анализом диаграммы рассеивания для этой пары переменных, где видна умеренная положительная корреляционная связь.

Таким образом, увеличение расстояния перелета действительно влияет на увеличение стоимости авиабилета, что является логичным следствием дополнительных затрат на дальние рейсы, такие как топливо, обслуживание и операционные расходы авиакомпаний.

**Сложная гипотеза 1:** В течение определенного количества дней с момента покупки до вылета стоимость авиабилетов увеличивается, после чего остается на том же уровне.

Гипотеза о положительном влиянии количества дней с момента покупки до вылета на стоимость авиабилета подтверждается в результате графического анализа и корреляционных коэффициентов. При этом, для сравнения прироста стоимости билетов необходим дополнительный анализ с построением регрессионных моделей, что будет рассмотрено ниже.

**Сложная гипотеза 2:** С увеличением расстояния, скорость роста цены авиабилета на внутренних рейсах ниже чем на зарубежных.

Гипотеза о положительном влиянии роста расстояния на стоимость авиабилета подтверждается в результате графического анализа и полученных коэффициентов корреляции. При этом, для сравнения скорости роста стоимости авиабилета на внутренних и внешних перелетах при увеличении расстояния необходим дополнительный анализ с построением регрессионных моделей, что будет рассмотрено ниже.

#### **3. Проверка гипотез с помощью моделирования**

##### 3.1. Построение базовой модели.

В качестве базовой модели используется регрессионная модель с линейным включением всех значимых переменных.

Выборка разделяется случайным образом на обучающую (80% от общего объема) для построения модели и тестовую (20%) для проверки прогностических свойств. Для составления базовой модели также сделаем из порядковых и номинальных переменных несколько dummy-переменных, которые будут являться индикаторами.

Таким образом, уравнение регрессии можно записать следующим способом:

price = a0 + a1 \* search\_weekday + a2 \* depart\_weekday + a3 \* number\_of\_changes\_category\_1 + a4 \* number\_of\_changes\_category\_2 + a5 \* number\_of\_changes\_category\_3 + a6 \* airline\_category\_Azimut + a7 \* airline\_category\_Nordwind + a8 \* airline\_category\_Other + a9 \* airline\_category\_Pegasus + a10 \* airline\_category\_Pobeda + a11 \* airline\_category\_S7 + a12 \* airline\_category\_Turkish + a13 \* airline\_category\_Uzbekistan + a14 \* destination\_category\_culture + a15 \* destination\_category\_domestic + a16 \* destination\_category\_tourism + a17 \* diff\_days + a18 \* distance

**Гипотезу 1** о том, что рост расстояния перелета города положительно влияет на цену, проверим следующим образом:

Если гипотеза 𝐻0 о том, что расстояние не влияет на стоимость, верна, то коэффициент при переменной distance a18 = 0. Если верна альтернативная гипотеза 𝐻1 о том, что возрастание расстояния влияет на стоимость в сторону ее увеличения, то 𝑎18 > 0.

**Для сложной гипотезы 1** введем переменную diff\_days\_squared = diff\_days\*\*2

Уравнение регрессии будет иметь вид:

price = a0 + a1 \* search\_weekday + a2 \* depart\_weekday + a3 \* number\_of\_changes\_category\_1 + a4 \* number\_of\_changes\_category\_2 + a5 \* number\_of\_changes\_category\_3 + a6 \* airline\_category\_Azimut + a7 \* airline\_category\_Nordwind + a8 \* airline\_category\_Other + a9 \* airline\_category\_Pegasus + a10 \* airline\_category\_Pobeda + a11 \* airline\_category\_S7 + a12 \* airline\_category\_Turkish + a13 \* airline\_category\_Uzbekistan + a14 \* destination\_category\_culture + a15 \* destination\_category\_domestic + a16 \* destination\_category\_tourism + a17 \* diff\_days + a18 \* distance + a19 \* diff\_days\_squared

Если коэффициент для diff\_days отрицательный (с уменьшением дней цена растет), а для diff\_days\_squared положительный и оба значимы, это указывает на то, что цена сначала увеличивается, а затем стабилизируется или уменьшается.

**Для сложной гипотезы 2** введем переменные "dist\_domestic"и "dist\_not\_domestic". Уравнение регрессии будет иметь следующий вид:

price = a0 + a1 \* search\_weekday + a2 \* depart\_weekday + a3 \* number\_of\_changes\_category\_1 + a4 \* number\_of\_changes\_category\_2 + a5 \* number\_of\_changes\_category\_3 + a6 \* airline\_category\_Azimut + a7 \* airline\_category\_Nordwind + a8 \* airline\_category\_Other + a9 \* airline\_category\_Pegasus + a10 \* airline\_category\_Pobeda + a11 \* airline\_category\_S7 + a12 \* airline\_category\_Turkish + a13 \* airline\_category\_Uzbekistan + a14 \* destination\_category\_culture + a15 \* destination\_category\_domestic + a16 \* destination\_category\_tourism + a17 \* diff\_days + a18\_0 \* dist\_domestic + a18\_1 \* dist\_not\_domestic

Причем, если гипотеза верна, то будет a18\_0 < a18\_1, a18\_0 > 0, a18\_1 > 0

##### 3.2. Анализ наличия гетероскедастичности

| LM Statistic | 7796.38 |
| --- | --- |
| P-value | 0 |
| F-statistic | 79.56 |
| P-value | 0 |

*Таблица 9. тест Уайта на наличие гетероскедастичности*

Была произведена проверка базовой модели на наличие гетероскедастичности. Результатом проверки стало отвержение нулевой гипотезы о гомоскедастичности модели (p-value близко к нулю)

##### 3.3. Проверка мультиколлинеарности

|  | изначально |
| --- | --- |
| const | 30.767966 |
| search\_weekday | 1.001149 |
| depart\_weekday | 1.007606 |
| number\_of\_changes\_category\_1 | 1.834520 |
| number\_of\_changes\_category\_2 | 2.512865 |
| number\_of\_changes\_category\_3 | 1.642167 |
| airline\_category\_Azimut | 1.991553 |
| airline\_category\_Nordwind | 1.520187 |
| airline\_category\_Other | 1.956152 |
| airline\_category\_Pegasus | 1.946393 |
| airline\_category\_Pobeda | 1.779132 |
| airline\_category\_S7 | 2.078393 |
| airline\_category\_Turkish | 2.361369 |
| airline\_category\_Uzbekistan | 2.403660 |
| destination\_category\_culture | 3.607605 |
| destination\_category\_domestic | 5.377524 |
| destination\_category\_tourism | 3.591784 |
| diff\_days | 1.188886 |
| distance | 1.403083 |

*Таблица 10. проверка на мультиколлинеарность*

Из таблицы видно, что на базовой модели некоторые значения VIF для столбцов destination\_category больше 3. Но эти переменные нужны нам для анализа одной из гипотез, так что их удалять мы не будем.

В случае константы это ожидаемый эффект, от переменной destination\_category\_domestic мы избавились. В случае остальных переменных значение VIF небольшое (VIF < 3), а значит мультиколлинеарности не наблюдается

##### 3.2. Проверка гипотез с помощью моделирования

Проверим выдвинутые нами гипотезы.

**Гипотеза 1:** посмотрим на коэффициент модели перед ‘distance’: a\_18 = 4611.14. Он положительный, значит мы отвергаем нулевую гипотезу о том, что расстояние не влияет на стоимость в пользу альтернативы. Получаем, что рост расстояния перелета положительно влияет на цену авиабилета.

**Гипотеза 2:** Воспользуемся модифицированной моделью, в которую мы добавили два новых признака. Для проверки гипотезы посмотрим на коэффициенты перед этими признаками: distance\_domestic и distance\_abroad.

a\_18\_0 (distance\_domestic) = 2889.29, a\_18\_1 (distance\_abroad) = 6342.84

Получили, что коэффициент, отвечающий за силу влияния расстояния на цену, если перелет внутренний, меньше, чем если перелет внешний. 2889.29 < 6342.84, а значит, заграничные перелеты действительно дорожают быстрее внутренних с увеличением расстояния.

**Гипотеза 3:** Воспользуемся второй модифицированной моделью, в которую мы добавим признак diff\_days\_squared и посмотрим на коэффициенты перед ним и перед diff\_days. a\_18\_0 (diff\_days) = -1680.8560886009666, a\_18\_1 (diff\_days\_squared) = 588.0721576463059

Полученные значения подтверждают, что с уменьшением дней до перелета, цена на билеты возрастает до какого-то момента, а потом рост цены замедляется и остается примерно на том же уровне

Соберем итоговую модель, в которую включим все признаки, необходимые для проверки всех гипотез.

Коэффициенты модели:

[ 0. 202.69522644 61.74539561 1812.85115607

4105.29863358 3606.39937969 -1402.57451738 -797.50075172

-892.02391943 324.9014432 -758.63890463 -275.14411785

3336.9985792 12.91711065 1390.5703039 899.72296031

1666.55851663 -1484.12315224 269.33278869 2902.49449323

6342.92142613]

Не значимыми переменными являются const и airline\_category\_Uzbekistan\_Airways. С ними будем работать при оптимизации модели.

##### 3.3. Оптимизация итоговой модели, сравнение качества моделей.

| Номер / Критерий | R^2 | Adj R^2 | Akaike |
| --- | --- | --- | --- |
| 1. Базовая модель | 0.726 | 0.726 | 5.562e+05 |
| 2. Модифицированная модель | 0.889 | 0.888 | 5.567e+05 |
| 3. Модифицированная модель Итоговая | 0.735 | 0.734 | 1.275e+04 |

*Таблица 11. Сравнение качества построенных моделей.*

Анализируя таблицу (15), можно заметить, что в соответствии с показателем Akaike (наибольшее значение) итоговая оптимизированная модель является лучшей. Это объясняется тем, что в ней мы учли как результаты проверки гипотез, так и результаты предварительного анализа данных. Данное качество модели было достигнуто путем удаления из модели незначимое фиктивное переменной: *“airline\_category\_Uzbekistan Airways”*

##### 3.4. Проверка прогностических способностей модели

| Номер\Критерий | Среднеквадратическая погрешность | Абсолютная  погрешность | Доверительная  вероятность |
| --- | --- | --- | --- |
| 1. Базовая модель | 7109.73 | 4674.21 | 0.013 |
| 2. Итоговая модель | 7148.61 | 4632.08 | 0.013 |

*Таблица 11. Сравнение прогностических способностей моделей*

Из таблицы можно увидеть, что среднеквадратичная погрешность немного увеличилась, а абсолютная уменьшилась. В среднем, ошибка сильно не изменилась.

#### **4. Заключение**

На основании построенной модели мы смогли подтвердить все три гипотезы.

1. Анализ показал, что с увеличением расстояния перелета, увеличивается и цена авиабилета. Это подтверждает гипотезу о положительном влиянии расстояния на цену.
2. При увеличении расстояния, цена быстрее растет на зарубежных перелетах, чем на внутренних. Действительно, цены на зарубежные авиабилеты могут быстрее расти из-за высоких себестоимости и налогов, валютных колебаний и меньшей конкуренции, чем на внутренних направлениях. Эти факторы влияют на базовую стоимость билетов и могут приводить к более быстрому росту цен на международные перелеты.
3. Цена на авиабилет сначала увеличивается при уменьшении количества дней до отлета, а потом стабилизируется или начинает падать. Эта гипотеза также подтвердилась. Это явление может объясняться стратегиями ценообразования авиакомпаний. В начальной фазе бронирования, когда времени до отлета остается немного, спрос на билеты часто возрастает, и авиакомпании могут повышать цены. После этого, когда остается меньше времени до вылета, важность цены для пассажиров возрастает, и авиакомпании могут корректировать цены в зависимости от спроса и заполненности рейса. Эта стратегия помогает авиакомпаниям максимизировать доходы от продажи билетов в различные периоды бронирования.