Задание-кейс.

В ходе выполнения задания были проанализированы данные о перелетах авиакомпании N за изучаемый период с 01.10.2021 по 30.09.2022. Основная задача исследования – проанализировать данные и построить наилучшую модель для оценки цены авиабилетов.

Для решения задания были поставлены следующие цели:

1. Скорректировать недостающие значения, удалить те данные, которые не пригодятся в последующем анализе;
2. Провести корреляционный анализ с выбранными факторами и проверить их значимость;
3. Провести регрессионный анализ, проверить значимость регрессии;
4. Построить итоговую модель и по ней рассчитать цену авиабилетов в конкретном примере;
5. Визуализировать данные, обосновать полученные результаты и дать личные рекомендации.

Как было описано выше, сначала были скорректированы первоначальные данные и удалены «лишние» переменные.

Поскольку основная наша цель – рассчитать цену авиабилета, мною были преобразованы некоторые переменные и были переведены в «бинарный формат», а именно:

- Переменная PAS_TYPE: если пассажир взрослый, то переменной присваивается значение 1, если пассажир ребенок – 2, если информации нет – 0.
- Переменная ROUTE_FLIGHT_TYPE: если перелет происходит внутри страны – переменной присваивается значение 1, если рейс международный – 0.
- Переменная FFP_FLAG: при наличии у пассажира программы лояльности – значение 1, если программы лояльности нет – 0.
- Переменная SALE_TYPE: при покупке билета онлайн – переменной присваивается значение 1, при покупке оффлайн – 0.

Помимо бинарных переменных была введена еще одна – разница в днях между датой покупки билета и датой вылета. Данная переменная позволяет в дальнейшем оценивать, насколько заранее пассажир купил себе авиабилет.

Некоторые переменные – координаты аэропортов были удалены (OC_lat, OC_long, DC_lat, DC_long), поскольку с помощью широты и долготы будет весьма проблематично

оценить рост стоимости билета (например, есть авиакомпании, которые базируются в определенном аэропорте и вылеты из крупных городов данной авиакомпанией выполняются только оттуда, или в целом есть авиакомпании, у которых выбор аэропорта в городе не влияет на стоимость, а также стоит учитывать, что в мире не так много городов с нескольким аэропортами по близости).

Вместо переменных о кодах городов и аэропортов было взято расстояние между городами рейса, так как именно длительность полета и расположение городов влияет на стоимость.

Итак, скорректировав все необходимые переменные, построим корреляционную таблицу (табл. 1)

Таблица 1 - Корреляционная матрица

	сумма покупки	разница между покупкой и вылетом	тип перелета	тип пассажира	программа лояльности	расстояние
сумма покупки	1					
разница между покупкой и вылетом	0,029465	1				
тип перелета	-0,10481	0,068227	1			
тип пассажира	0,013758	0,0917	-0,01296	1		
программа лояльности	-0,0335	0,022139	0,158025	-0,11294	1	
расстояние	0,59285	0,136256	-0,05776	0,026094	0,013996	1

Исходя из таблицы видно, что корреляция между выбранными факторами не очень сильная и мультиколлинеарность не наблюдается. Поэтому далее коэффициенты корреляции были проверены по критерию Стьюдента (табл. 2):

Таблица 2 – Рассчитанные значения для проверки значимости коэффициентов

	сумма покупки	разница между покупкой и вылетом	тип перелета	тип пассажира	программа лояльности
разница между покупкой и вылетом	6,56				
тип перелета	-23,45	15,22			
тип пассажира	3,06	20,49	-2,88		
программа лояльности	-7,46	4,93	35,62	-25,30	
расстояние	163,84	30,61	-12,88	5,81	3,12

Цветом выделены те коэффициенты, которые являются значимыми, поскольку превышают критическое значение (=1,96).

Далее был выполнен регрессионный анализ и были получены следующие значения (рис. 1):

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,599638415
R-квадрат	0,359566229
Нормированный R-квадрат	0,359502174
Стандартная ошибка	174,4201567
Наблюдения	49997

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	5	853866231,5	170773246,3	5613,406456	0
Остаток	49991	1520845751	30422,39105		
Итого	49996	2374711982			

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*	*ижние 95%*	*ерхние 95%*	*жние 95,0%*	*рхние 95,0%*
Y-пересечение	286,6431914	4,363385617	65,6928396	0	278,0909	295,1955	278,0909	295,1955
разница между покупкой и вылетом	-0,330023493	0,02574672	-12,81807884	1,48851E-37	-0,38049	-0,27956	-0,38049	-0,27956
тип перелета	-50,41532428	2,947357295	-17,10526388	2,083E-65	-56,1922	-44,6385	-56,1922	-44,6385
тип пассажира	-1,466402253	2,873594831	-0,51030237	0,609841896	-7,09868	4,165876	-7,09868	4,165876
программа лояльности	-13,9440002	1,632347756	-8,542297524	1,35224E-17	-17,1434	-10,7446	-17,1434	-10,7446
расстояние	0,098895998	0,000601012	164,5491144	0	0,097718	0,100074	0,097718	0,100074

Рисунок 1 – Вывод результатов по построенной регрессионной модели

Полученная регрессионная модель была проверена на значимость с помощью критерия Фишера. Критическое значение равно 2,214, что значительно меньше наблюдаемого значения, поэтому можно подтвердить гипотезу о том, что регрессия значима. Также видим, что значение R-квадрата неплохое – около 0,6 и все переменные являются значимыми, поскольку их значение p-value меньше 0,05.

Исходя из полученных результатов получим следующую регрессионную модель:
Y=286.6-0.33*x1-50.42*x2-1.46*x3-13.94*x4+0.09*x5.

Далее, используя построенную модель была рассчитана стоимость авиабилетов для определенных ситуаций (выполним 7 пункт задания):

а) Авиабилеты для гражданина Иванова И.И. будут стоить 430,243 у.е. туда и 428,59 у.е. обратно.

Исходя из модели, фактор x1 в билете туда будет равен 7 (за столько дней заранее Иванов И.И. покупает билет), фактор x2=1 поскольку рейс в пределах России, фактор x3=1 – т.е. взрослый человек, x4=0 (поскольку мы не знаем, есть ли у него программа лояльности) и фактор x5=2000 (возьмем условное расстояние между городами).

Для билета обратно расчеты будут примерно те же, за исключением фактора x1, он будет равен 12, потому что билет обратно покупается еще более заранее, чем билет туда.

Б) Авиабилеты для гражданина Иванова И.И. и его сына будут стоить: 200,686 у.е. для Иванова И.И. туда, 192,11 у.е. для гражданина Иванова И.И. обратно, 199,22 у.е. для сына туда и 190,64 у.е. для сына обратно.

Модель будет рассчитываться так же, как и в прошлом примере, но у данного случая дни (фактор x1) будут равны 61 день туда и 87 дней обратно, у ребенка будет тип пассажира (фактор x2) равен 2 и в обоих билетах программа лояльности (фактор x4) будет иметь значение 1, поскольку у гражданина есть накопленные бонусы. (в факторе x5 расстояние также бралось условно = 1000 км)

Исходя из примеров, мы видим, что модель грамотно показывает, что для ребенка билет стоит дешевле, при программе лояльности также цена несколько снижается, при внутренних рейсах цена также ниже. И немаловажным фактом является то, что при покупке билета заранее – цена также снижается.

Таким образом, была выполнена работа по оценке предварительных цен на авиабилеты. Были отобраны необходимые факторы, проведен корреляционный и регрессионный анализ, коэффициенты корреляции и регрессионная модель были проверены на адекватность. На примере удалось отметить, что модель действительно работает и грамотно рассчитывает цену билета исходя из заданных ранее параметров и моментов, на которые делался упор.

Хотелось отметить немаловажный момент, на который можно обратить внимание при последующем анализе – сезонность, которая довольно часто и сильно влияет на цену авиабилетов. К сожалению, в силу небольшого количества времени не удалось построить регрессионную модель с учетом сезонов отпусков и других праздников, поэтому был построен небольшой дашборд, на котором визуально можно увидеть сезонность и оценить ее значимость.

Дашборд доступен по ссылке ниже: https://datalens.yandex/or712sv9ul5g8