Проект по эконометрике

Построение модели ценообразования номеров в гостиницах Сочи для одного человека

Проект выполнили: Горюнова Екатерина, Павлек Екатерина, Тюльберова Анна

Задача	Кто выполнял
Парсинг данных	Тюльберова Анна
Визуализация данных	Горюнова Екатерина (гистограммы, точечные диаграммы) Павлек Екатерина (облака слов)
Обработка данных	Горюнова, Павлек, Тюльберова
Построение моделей	Павлек (линейная и полулогарифмическая модели), Горюнова (линейная в логарифах) Тюльберова (дополнительная модель)
Проведение тестов	Павлек (тесты на значимость переменных, тест Бокса-Кокса) Горюнова (значимость модели в целом)
Проверка мультиколлинеарности	Павлек (построение VIF) Тюльберова (создание новых переменных) Горюнова (построение моделей по исправленным данным)
Проверка гетероскедастичности	Павлек (тесты Уайта и Бройша-Пагана) Горюнова (построение моделей с поправками Уайта)
Анализ выбросов	Тюльберова (проведение теста DFFITS)
Спецификация модели	Павлек (тест Рамсея)
Оформление итогового текстового файлы	Горюнова Павлек Тюльберова (доп. модель)

Итоговый вклад

Горюнова Екатерина	33.3%
Павлек Екатерина	33.3%
Тюльберова Анна	33.3%
Коты	0.1%

Введение:

В качестве данных нами были взяты минимальные цены на отели в Сочи с сайта Островок (https://ostrovok.ru/hotel/russia/sochi/). В качестве зависимой переменной выступает цена одного номера для одного человека (цена взять минимальная по отелю, она не зависит ни от времени посещения отеля, ни от загруженности, в данных отсутствует сезонность). После удаления всех пропусков и фильтрации отелей по местоположению, в выборке осталось 1808 наблюдений.

Характеристиками товара являются:

'web' - ссылка на отель

'perks' - главные удобства отеля, представляет собой список различных услуг

'amenities' - дополнительные услуги

'dist_sea' - расстояние до моря в метрах

'dist_city' - расстояние до центра в метрах

'stars' - звездность (от 0 до 5)

'feedb' - средняя оценка на основе отзывов

'feedb_n' - количество отзывов

'num_rooms' - количество номеров в отеле

'kids' - количество услуг для детей (детские площадки, возможность нанять няню и т. д.)

'location' - расположение отеля, изначально в выборку попали отели из несколько городов недалеко от Сочи, в ходе работы были оставлены отели в самом городе.

Для удобства мы создали дополнительные переменные, отвечающие за количество основных удобств (perks) и дополнительных удобств (amenities): 'perks_n', 'amenities_n'

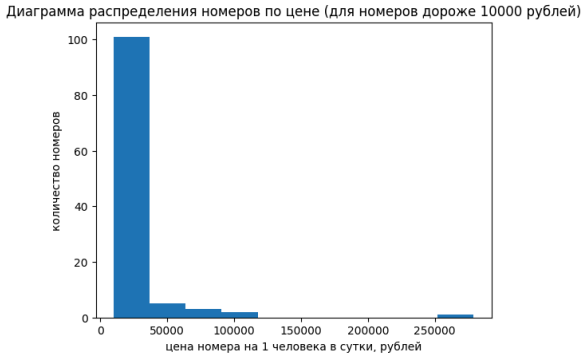
А также несколько бинарных переменных, отвечающих за наличие самых важных услуг:

'Restaurant_bin', 'Pool_bin', 'Internet_bin', 'Jacuzzi_bin', 'Disaibaled_bin', 'Conditioner_bin', 'Conference_room_bin', 'Kitchen_bin', 'Smoking_available_bin', 'Parking_bin', 'Beach_nearby_bin', 'Kids_bin', 'Pets_bin', 'SPA_bin', 'Transfer_bin', 'Fitness_bin', 'tv_bin', 'brekf_bin', 'fridge_bin', 'heat_bin', 'iron_bin', 'smokefree_bin', 'invalid_bin'.

Мы также построили распределения основных переменных и привели их ниже.

Данные распределены неравномерно: количество отелей 0 звезд существенно выше остальных (звездности которых не указаны), однако судя по остальным переменным, качество этих отелей существенно различается

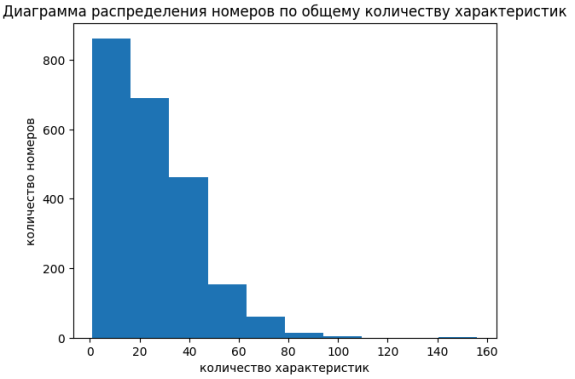
Цены на отели в основном лежат в диапазоне до 10000, примерно 100 отелей в диапазоне от 10 до 25 тысяч и небольшая часть превышает 50 тысяч.



Большинство отелей расположено довольно близко к пляжу, тогда как относительно центра города они поделены примерно на 3 кластера: расстояние меньше 10 км, от 10 до 30 км и от 40 до 60 км. Это может быть связано с тем, что пляжный курорт растянут вдоль моря, поэтому расстояние до центра существенно отличается.

Также мы построили облака слов для переменных perks и amenities, которые показали, что наиболее популярными удобствами является расстояние до пляжа, бесплатная парковка, интернет и удобства для детей, а из дополнительных услуг наиболее часто встречаются гладильные принадлежности, и техника (стиральная машина, утюг, кондиционер и отопление, кухня, холодильник и телевизор).

Количество удобств практически во всех отелях равно 5, тогда как количество дополнительных услуг примерно для половины отелей составляет от 0 до 20, а для большинства отелей этот показатель не превышает 40.



Для построения моделей мы оставили только отели, расположенные непосредственно в Сочи, поскольку в этой группе находилось наибольшее количество отелей (1808 отелей без учета пропусков в данных). Другие отели (например, в Адлере или в поселениях вокруг Сочи) были убраны из выборки для того, чтобы она была более однородной.

Построение моделей:

Мы построили линейную, полулогарифмическую и логарифмическую модели для того, чтобы сравнить их и в дальнейшем использовать модель с наиболее высоким качеством.

Для линейной модели значение R^2 оказалось равно 0.155, а скорректированного R^2 0.142. Значимыми переменными оказались переменные: dist_city, stars, feedb, feedb_n, perks_n, amenities_n, Restaurant_bin, Internet_bin, Kids_bin, Pets_bin, Fitness_bin. Результаты данной модели и коэффициенты переменных представлены в “Таблице 1” итогового файла с кодом.

Для полулогарифмической модели значение R^2 равно 0.246, а скорректированного R^2 0.234. Значимыми переменными для данной модели оказались переменные: dist_city, stars, feedb, feedb_n, amenities_n, Pool_bin, Internet_bin, Disaibaled_bin, Fitness_bin, tv_bin. Результаты данной модели и коэффициенты переменных представлены в “Таблице 2” итогового файла с кодом.

Для создания линейной в логарифмах модели мы использовали логарифмы количественных переменных: dist_city, num_rooms, perks_n, amenities_n, dist_sea, feedb_n. R^2 данной модели оказался равен 0.249, а скорректированный R^2 0.237, а значимыми переменными: stars, Pool_bin, Internet_bin, Disaibaled_bin, Beach_nearby_bin, Fitness_bin, brekf_bin, dist_city, num_rooms, amenities_n, dist_sea, feedb_n. Результаты данной модели и коэффициенты переменных представлены в “Таблице 3” итогового файла с кодом.

Сравнение моделей:

Для сравнения полулогарифмической и линейной моделей мы использовали тест Бокса-Кокса с преобразованием Зарембки. Значение тестовой статистики, которая имеет хи-квадрат распределение с 1 степенью свободы, оказалось равно 1701.68, что существенно превышает критическое значение, равное 3.84 на уровне значимости 0.05. Таким образом, на данном уровне значимости нулевая гипотеза об одинаковых уровнях подгонки линейной и полулогарифмической моделей отвергается и лучше модель с более низким RSS. Поскольку для полулогарифмической модели RSS значительно ниже, чем для линейной, из этих двух моделей более качественной является

полулогарифмическая модель. Результаты RSS и расчет тестовых статистик представлены в блоке “сравнение линейной и полулогарифмической моделей” итогового файла с кодом.
Для выбора между полулогарифмической и логарифмической моделями мы сравнили RSS этих моделей и получили, что логарифмическая модель лучше полулогарифмической. Таким образом, мы использовали логарифмическую модель для дальнейших тестов. Расчет показателей и их значения представлены в блоке “Итог сравнения моделей”.

Проверка мультиколлинеарности:
Также мы провели проверку данных на мультиколлинеарность с помощью построения корреляционной матрицы и VIF для каждой из переменных. В результате первичного анализа было выявлено, что корреляция между признаками не слишком высокая (не превышала 0.74), однако VIF переменных Internet_bin, dist_city, perks_n, amenities_n была значительно выше 10, что указывает на наличие мультиколлинеарности. Кроме того, показатель переменной dist_sea оказался равен 9.91. Для того, чтобы избавиться от этой проблемы, мы создали две новые переменные: amen_only, которая отвечает за дополнительные услуги отеля (те, которых нет в общих услугах (perks), но есть в доп. услугах (amenities), так как perks может дублировать amenities) и sum_dist_w, которая равна логарифму суммы расстояний до моря и до города. Мы также провели оптимизацию по весам для суммы расстояний, однако для всех соотношений коэффициентов показатели VIF и R^2 модели оказались практически одинаковыми, поэтому мы приняли решение использовать равные коэффициенты, равные 1. Очевидно, что чем больше расстояние, тем дальше отель находится от моря, центра или обоих, значит его цена, скорее всего, будет ниже. В результате данных преобразований значения VIF для всех переменных оказался меньше 10, то есть проблема мультиколлинеарности была решена. Мы оценили модель по новым данным, R^2 стал равен 0.262, а значимые переменные - stars, feedb, Pool_bin, Internet_bin, Disaibaled_bin, Fitness_bin, tv_bin, fridge_bin, num_rooms, amen_only, sum_dist_w.

Проверка гетероскедастичности и выбросов:

Кроме того, мы провели проверку модели на гетероскедастичность. Для этого мы использовали тесты Уайта и Бройша-Пагана. Для обоих этих тестов нулевая гипотеза состоит в отсутствии в данных гетероскедастичности. Сделав тест Уайта по модели, которая получилась после устранения мультиколлинеарности, мы получили, что p-value примерно равен 0.0001, тест Бройша-Пагана также показал, что p-value очень близко к нулю, поэтому мы должны отвергнуть нулевую гипотезу на любом разумном уровне значимости и признать наличие гетероскедастичности в данных. Для того, чтобы избавиться от гетероскедастичности в данных, мы использовали поправки Уайта HC3. Также мы проверили модель на выбросы с помощью встроенного метода DFFITS. В результате проверки выбросами оказались 64 наблюдения. Мы проанализировали значения переменных и пришли к тому, что эти наблюдения не являются влиятельными, так как среди них не выделялись отдельные группы переменных (большинство выбросов имели звездность 0, различные расстояния до центра и моря, разное количество услуг и отзывов и т.д.), поэтому мы приняли решение удалить их из выборки.

Проверка спецификации модели:

Кроме того, мы провели тест Рамсея на функциональную форму модели. Нулевая гипотеза данного теста заключается в том, что выбранная спецификация модели является верной. Мы проводили тест Рамсея для итоговой модели (то есть для линейной в логарифмах модели без выбросов, мультиколлинеарности и гетероскедастичности), добавляя в модель квадрат предсказанных значений. Наблюдаемое значение F-статистики оказалось больше критического для уровня значимости равного 0.05 (наблюдаемое значение равно 7.03, а критическое 3.8. Таким образом, тест Рамсея показал, что спецификация модели неверна. Для исправления ситуации мы добавили в модель квадраты всех переменных, кроме тех, которые являются бинарными. Результаты модели представлены в таблице 5 итогового файла в разделе “Тест Рамсея”. Получившийся R^2 равен 0.373, что существенно выше, чем в предыдущих моделях. После этого мы провели повторный тест Рамсея (результаты оцененной регрессии в таблице 6) и в этот раз наблюдаемое значение F-статистики оказалось ниже критического, поэтому у нас нет

оснований для отвержения нулевой гипотезы в повторном тесте Рамсея и мы можем говорить о том, что спецификация модели с квадратами переменных подходит лучше, чем без них. Кроме того, мы оценили модель. добавив в нее кубы переменных для того, чтобы проверить, не улучшит ли это модель. Результаты данной модели представлены в таблице 7. Таким образом, нами была выбрана модель с квадратами переменных, которые не являются дамми. Значимыми переменными оказались переменные stars, feedb, Restaurant_bin, Pool_bin, Internet_bin, Disaibaled_bin, brekf_bin, num_rooms, stars_^2, feedb_^2, num_rooms_^2.

Интерпретация итоговой модели:

Мы построили регрессию на переменных, которые оказались значимыми после проведения теста Рамсея.

$$ln(price) = 8.5954 - 0.1120 * stars - 0.0336 * feedb + 0.0775 * restaurant_{bin} + 0.0541 * pool_{bin} - 0.0340 * Internet_{bin} + 0.1497 * disabled_{bin} + 0.0729 * breakf_{bin} - 0.2681 * ln(num.rooms) + 0.0740 * stars^2 + 0.0035 * feedb^2 + 0.0413 * ln(num.rooms)^2$$

Таким образом, можно говорить о том, что зависимость цены номера от количества звезд, оценки отеля и логарифма количества доступных номеров нелинейна. При этом, поскольку коэффициенты при всех трех переменных больше 0, стоит говорить о том, что существует минимум влияния данных переменных на цену номера. Цена номеров в отелях, в которых есть ресторан в среднем на 0.0775% выше, чем в тех отелях, в которых нет ресторана, цена номеров в отелях, в которых есть бассейн в среднем на 0.0541% выше, чем в тех, в которых нет бассейна. Кроме того, доступность отеля для людей с ограниченными возможностями повышает цену номера в среднем на 0.1497%, а наличие услуги завтрака в отеле - на 0.0729%. Однако наличие бесплатного интернета в отеле в среднем снижает цену номера на 0.034%.

Модель множественного выбора:

На собранных нами данных была также построена модель множественного выбора. Мы выдвигаем гипотезу о том, что ценовую категорию отеля можно предсказать на основе его характеристик (таких как количество звезд, рейтинг, расстояние до моря и центра, наличие различных удобств и т. д.). То есть мы предполагаем, что существует связь между характеристиками отеля и его

категорией, и для проверки гипотезы мы построили мультиномиальную логит-модель.

Мы выделили следующие ценовые категории (по частоте встречающихся значений):

- “Бюджетные” отели (цена за ночь от 0 до 3362 рублей);
- “Средние” отели (цена за ночь от 3362 до 5500 рублей);
- “Дорогие” отели (цена за ночь более 5500 рублей).

Опишем основные показатели построенной модели. Значение логарифма функции правдоподобия равно -1642.8. Псевдо-R^2 (МакФаддена) равен 0.1526. Результаты модели показали, что значимыми (p-value < 0.05) оказались следующие переменные: 'dist_city', 'feedb', ‘feedb_n’, 'stars', 'Restaurant_bin', 'Pool_bin', 'Disaibaled_bin'. После того, как в модели были оставлены только значимые переменные, логарифм функции правдоподобия стал равен -1680.8, псевдо-R^2 0.133.

Для тестирования значимости уравнения в целом нам нужно проверить гипотезу о том, что коэффициенты при всех переменных равны нулю. Был проведен LR-тест. Нулевая гипотеза состоит в том, что уравнение незначимо (коэффициенты при переменных равны нулю). Тестовое значение LR-статистики равно 515.53, критическое значение хи-квадрат статистики равно 26.3. Так как тестовое значение больше критического, то на уровне значимости 5% мы можем отвергнуть нулевую гипотезу. P-value близко к нулю. Следовательно, уравнение является значимым. Это значит, что действительно прослеживается зависимость ценовой категории отеля от таких характеристик как: расстояние до города, рейтинг по отзывам, количество звезд, наличие различных удобств (отопление, телевизор, кондиционер) и т. д. Гипотеза о зависимости не отвергается.

Теперь перейдем к интерпретации. Найдем предельный эффект изменения зависимой переменной (посчитаем производную вероятностей по каждому признаку).

При увеличении числа звезд на одну вероятность попадания отеля в отдельную категорию увеличится на 17.9 п. п. для “средней” и на 7.5 п. п. для “дорогой” категорий.

Наличие бара или ресторана увеличивает вероятность попадания отеля в “среднюю” категорию на 27.7 п. п. и в “дорогую” категорию на 26.1 п. п.

Увеличение среднего отзыва на 1 пункт уменьшает вероятность попадания отеля в “дорогую” категорию на 93.3 п. п. (не является корректной).
Увеличение расстояния до центра города на 1 метр уменьшает вероятность попадания отеля в “дорогую” категорию на 2.45%.
Наличие бассейна увеличивает вероятность попадания отеля в “дорогую” категорию на 35 п. п.
Наличие удобств для людей с ограниченными возможностями увеличивает вероятность попадания отеля в “дорогую” категорию на 42.9 п. п.
Данная категоризация цен отеля позволяет оптимизировать поиск наиболее подходящего варианта для конкретного человека. С помощью характеристик можно заранее определить, к какой категории относится тот или иной отель. Даже в сезон или при других условиях, когда цены на проживание могут сильно колебаться, для потенциального гостя в любом случае будет известен тип отеля.

Выводы:

В рамках данной работы мы собрали данные по отелям города Сочи, обработали их, проанализировали и на их основе построили несколько моделей и оценили гедонистическую ценовую функцию.
В качестве итоговой модели была выбрана линейная модель в логарифмах, так как она давала наилучшие результаты. Также в ходе анализа мы выявили мультиколлинеарность и гетероскедастичность в данных и приняли соответствующие меры по их устранению(созданием новых переменных на основе старых и применение поправок Уайта). Кроме того, мы проанализировали выбросы в модели и удалили их из модели, так как они не являлись влиятельными и только ухудшали качество. По результата теста Рамсея было получено, что в исходной модели пропущен квадрат количественных переменных, после добавления которых качество модели существенно улучшилось. В результате сильнее всего на цену отеля влияла переменная доступность отеля для людей с ограниченными возможностями.
В качестве дополнительной модели мы построили мультиномиальную логит-модель, с помощью которой проверили гипотезу о связи между ценовой категорией отеля и его характеристиками. По результатам модели данная гипотеза не была отвергнута.