Князева, Горячева,
Пластинина, Николина

Тема: Исследование зависимостей цены горнолыжных курортов мира от различных факторов

1. Сбор данных

Для выполнения анализа был использован набор данных с платформы Kaggle, содержащий следующие числовые переменные:

- Price — стоимость отдыха (в евро),
- Highest point — высота самой высокой точки курорта (в м),
- Beginner slopes — длина трасс для новичков (в км),
- Intermediate slopes — длина трасс средней сложности (в км),
- Difficult slopes — длина сложных трасс (в км),
- Total slopes — общая длина трасс (в км),
- Longest run — длина самой длинной трассы на курорте (в км),
- Snow cannons — количество снежных пушек,
- Surface lifts — количество бугельных подъемников,
- Chair lifts — количество кресельных подъемников,
- Gondola lifts — количество гондольных подъемников,
- Total lifts — общее количество подъемников на курорте,
- Lift capacity — суммарная пропускная способность подъемников (чел/час),
- Child friendly — подходит ли курорт для детей,
- Snowparks — наличие сноупарков,
- Nightskiing — возможность катания на освещенных склонах
- Season — нормальное начало и конец сезона на курорте

Мы исследуем цены в разных странах, было принято решение учитывать ВВП страны, так как ВВП не только отражает экономическую активность, но и влияет на ценообразование через динамику спроса, предложение и инфляционные процессы. Данные о ВВП мы также отдельно нашли на платформе Kaggle:

- GDP — ВВП региона, где находится курорт (в условных единицах)

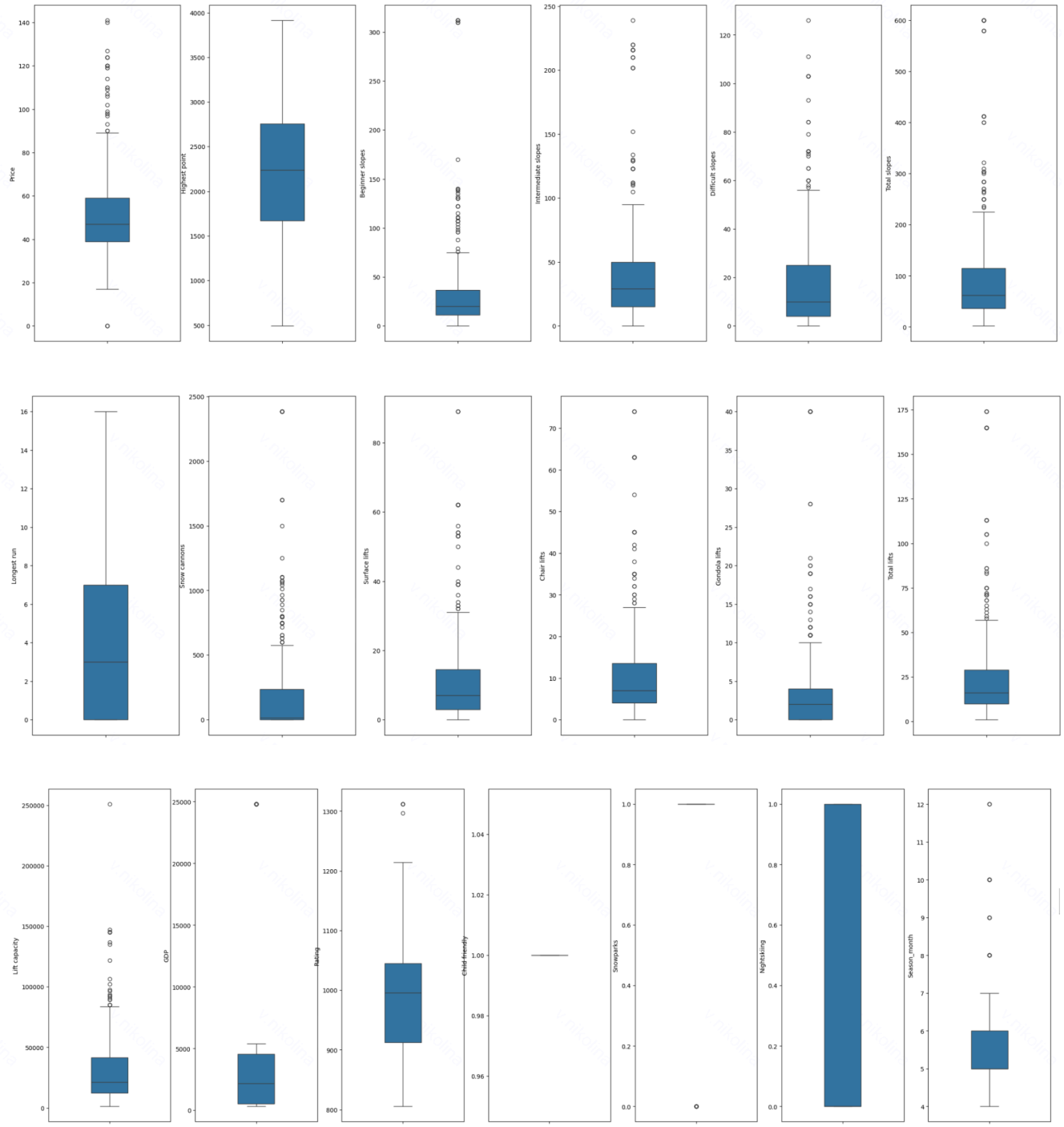
Также мы решили добавить такой фактор, как популярность курорта, возможно это будет влиять на спрос, а соответственно и на цену.

- Rating — рейтинг курорта (в баллах)

2. Предобработка данных

Необходимо преобразовать категориальные признаки в исчисляемые, в нашем случае это факторы Child friendly, Snowparks, Nightskiing, Season. Данные Season были неудобны для нас и мы преобразовали этот признак в количество месяцев в году, в которые работает курорт.

Анализируем выбросы собранных данных с помощью графиков Box plot.

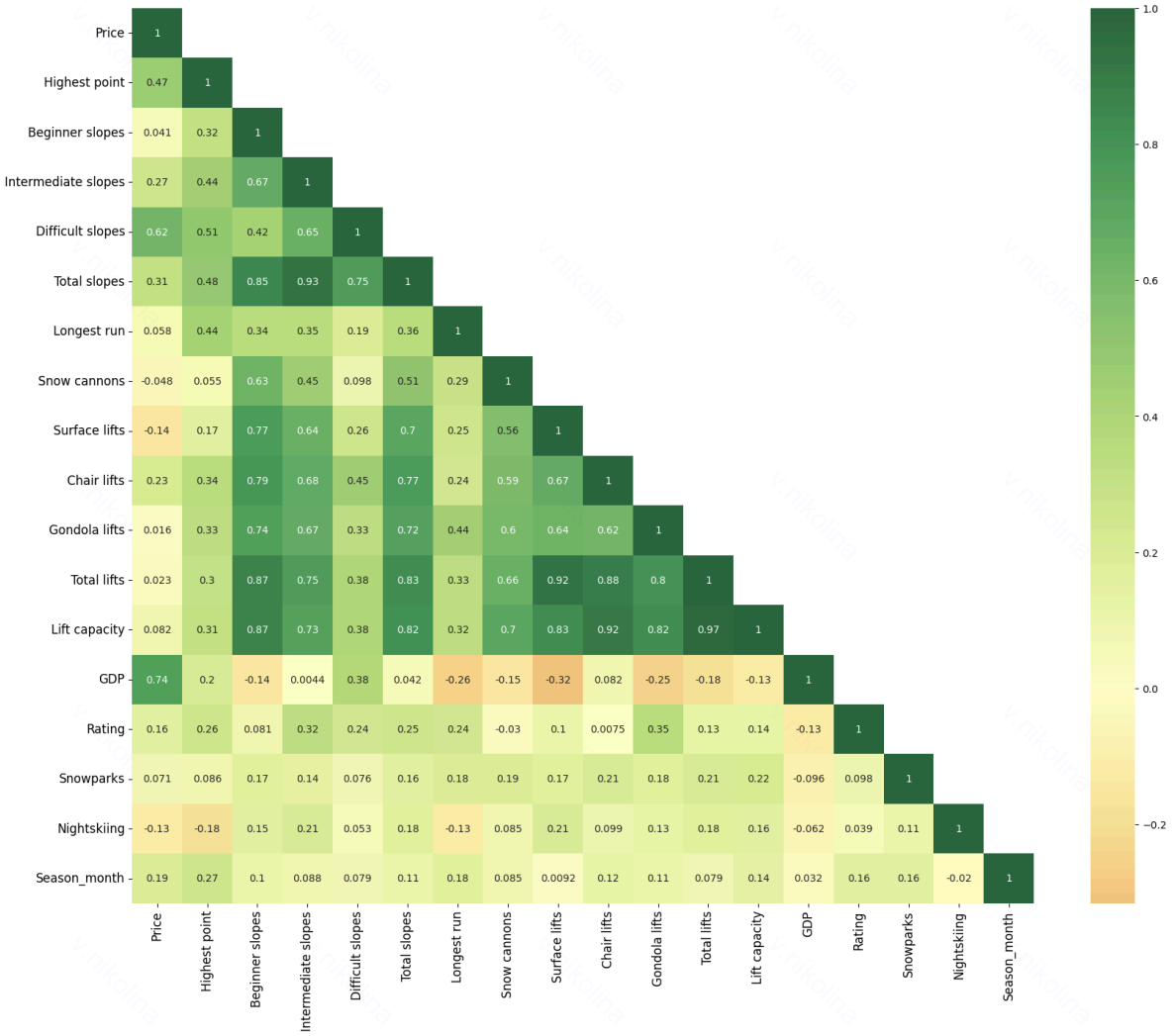


Отбрасываем выбросы, выходящие за пределы усов ящика. Также выставляем ограничения на факторы, основываясь на реальных величинах, например, суммарная пропускная способность не может быть отрицательной и т.д. Убедились в реальности собранных данных, проверив их на нескольких конкретных странах вручную.

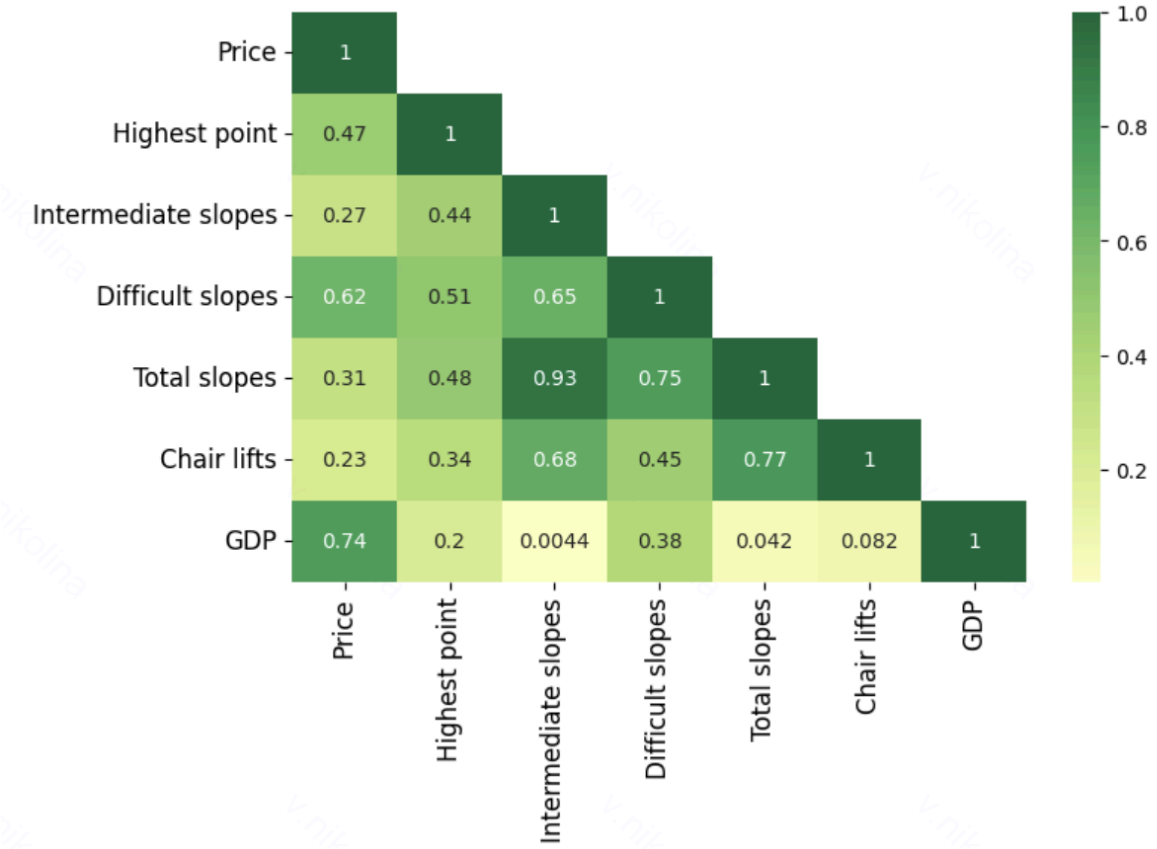
Из 499 курортов мира (изначальный объем данных) работаем с 211 данными после предобработки (из которых на 80% (168 курортов) - объем обучающих данных, 20% (43 курорта) - объем тестовых данных.

3. Построение матрицы корреляций

Получаем матрицу корреляций по 18 признакам и таргетной переменной - цена.



Смотрим, как факторы коррелируют между собой. Оставляем факторы с высокой парной корреляцией так, чтобы они не повторялись. Убираем признаки, слабые по шкале Чеддока (коэффициент корреляции < 0.2). В итоге имеем факторы: Highest point, Intermediate slopes, Difficult slopes, Total slopes, Chair lifts, GDP.



Необходимо выбрать факторы из коллинеарных (коэффициент корреляции >= 0.7), для этого используем частные коэффициенты корреляции.

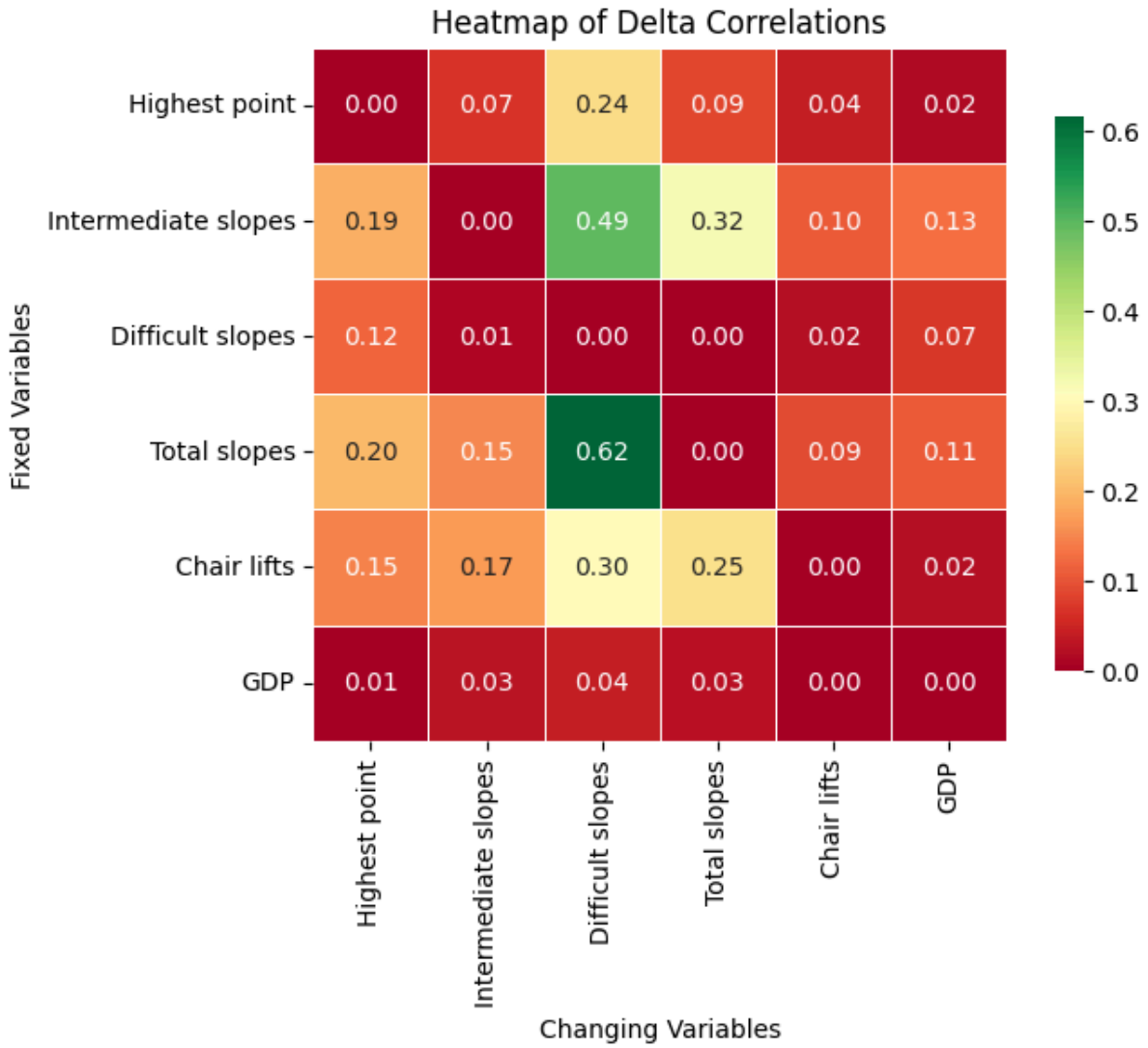
4. Отбор факторов посредством частных коэффициентов корреляции

Коэффициент частной корреляции z с x, если y закреплен в модели определяется следующим образом:

$$[r_{zx/y} = \frac{r_{zx} - r_{xy} \cdot r_{zy}}{\sqrt{(1 - r_{zx}^2)(1 - r_{xy}^2)}}]$$

Для каждого фактора закрепляем его в модели и перебираем оставшиеся факторы, считая ЧКК. Выясняем, какие факторы оказывает слабое влияние на изменение корреляции при их включении в модель. Чем меньше $r_{zx/y} - r_{zx}$, тем меньше влияния на модель оказывает фактор y.

Для удобной визуализации получаем матрицу дельт частных коэффициентов каждого фактора с каждым:



Интерпретацию матрицы дельт рассмотрим на примере. Смотрим по строчкам на Highest point, по столбцам на Difficult slopes. На их пересечении стоит 0.24. Это означает, что при закреплении Highest point, частный коэффициент корреляции между Difficult slopes и выходным параметром Price отличается по модулю от неочищенной корреляции между Difficult slopes и Price в матрице корреляций (он был равен 0.47) на 0.24 (назовем это дельтой).

Назовем “хорошими” те факторы, которые в строке имеют большое количество больших дельт.

Отбор факторов.

Total slopes. Все slopes коллинеарны и имеют большой коэффициент корреляции между собой. Выбираем фактор Total slopes, так как r=0.31 > r=0.27 у Intermediate slopes по матрице корреляций. Difficult slopes имеют слишком небольшую дельту. Chair lifts коллинеарен с Total slopes, а выбирая между ними, “хороший фактор” - Total slopes. Chair lifts не берем в модель.

Highest point. Фактор имеет дельту = 0.24, r=0.27, что достаточно высоко относительно остальных результатов.

GDP. r=0.74 в матрице корреляций - высокое значение. Несмотря на то, что при закреплении GDP в модели, фактор показался не хорошим, решили включить в модель и проверить значимость.

Включаем в модель следующие факторы: Highest point, Total slopes, GDP

5. Построение модели линейной регрессии

Модель имеет следующий вид: $y = b_0 + b_1 * x1 + b_2 * x2 + b_3 * x3$

Коэффициенты модели линейной регрессии вычисляем методом наименьших квадратов:

$$[\hat{B} = (X^T X)^{-1} X^T Z]$$

$$y = 57.27 + 3.35 * x_1 + 15.98 * x_2 + 6.53 * x_3$$

b0: 57.27
b1: 3.35 - Total slopes
b2: 15.98 - GDP
b3: 6.53 - Highest point
Считаем коэффициент детерминации модели: R^2 = 0.69, где

$$[R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}]$$

6. Проверка условий Гаусса-Маркова

Остаток e_i для i-го наблюдения определяется как: $[e_i = y_i - \hat{y}_i]$ [e_i = y_i - \hat{y}_i] где:

- y_i— фактическое значение целевой переменной,
- $\hat{y}_i$ — предсказанное значение целевой переменной на основе модели.

Условия Гаусса-Маркова для случайных ошибок:

- наличие гомоскедастичности - постоянства дисперсии остатков
- нормальное распределение остатков

График зависимости остатков от предсказанных значений. По графику видно, что остатки случайно разбросаны по обе стороны от нуля без какой-либо структуры. Следовательно, гомоскедастичность выполняется:

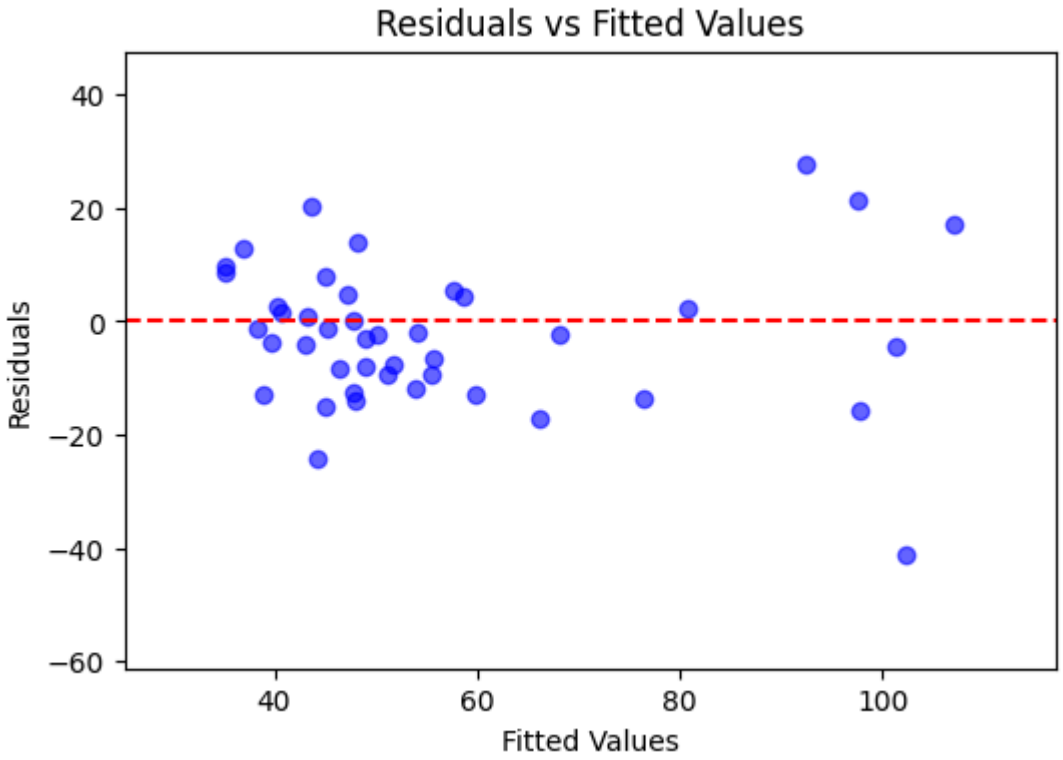


График распределения значения остатков для проверки остатков на нормальность. Видно, что нормальность выполняется:

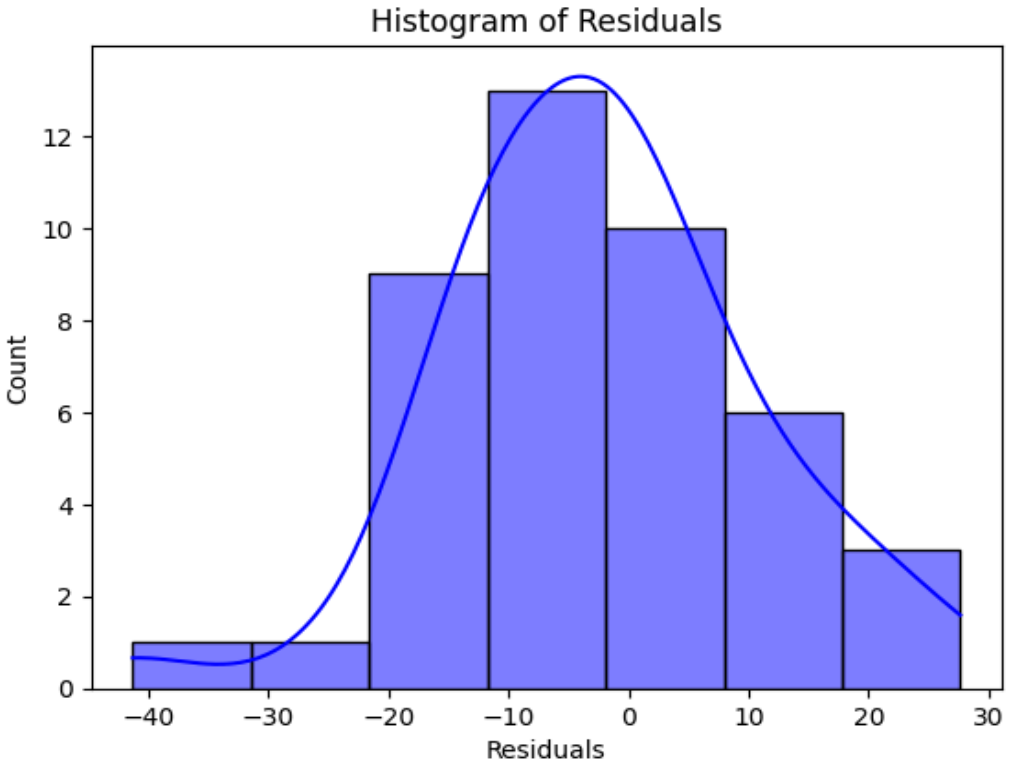
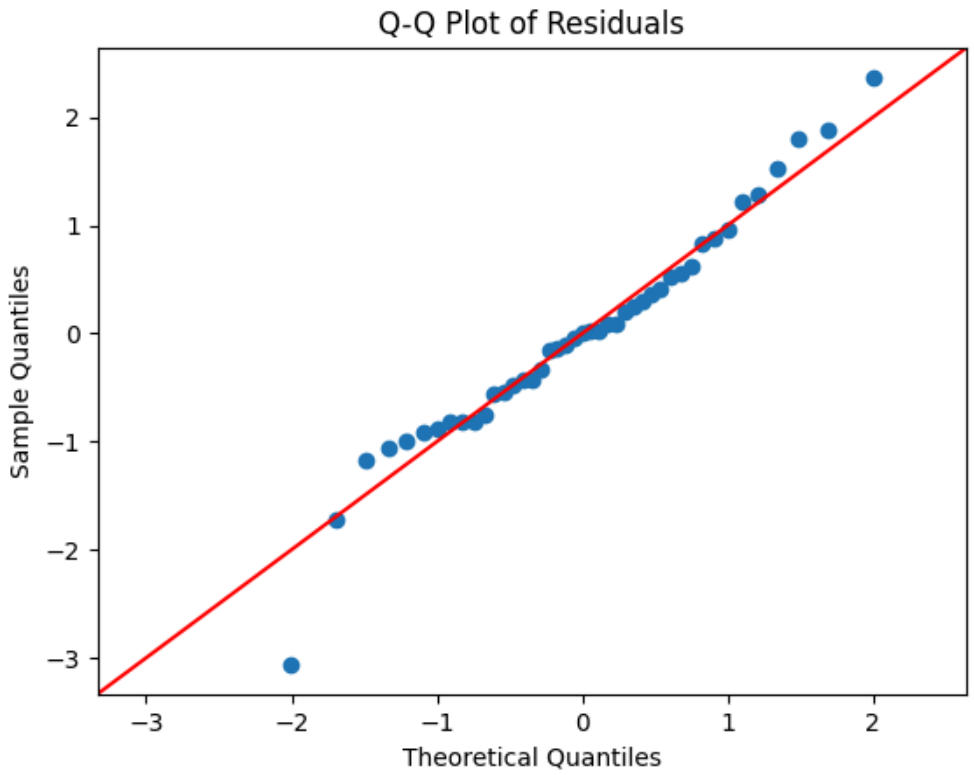


График квантилей - точки практически лежат на прямой, следовательно, остатки распределены нормально.



Условия Гаусса-Маркова выполняются. Модель в имеет право на существование.

7. Проверка значимости модели в целом

Постановка гипотез

- Нулевая гипотеза (H_0): $R^2 = 0$
- Альтернативная гипотеза (H_1): $R^2 \neq 0$

Уровень значимости 0.05

Формула F-статистики:

$$[F = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}}]$$

F - расчетная: 45.36

F - критическая=F(0.05, p-1, n-p): 3.23

F-расчетная > F-критическая. Отвергаем H_0 в пользу H_1. Модель значима на уровне значимости 0.05.

8. Проверка значимости коэффициентов

Постановка гипотез

- Нулевая гипотеза (H_0): $b_i = b_{0i} = 0$
- Альтернативная гипотеза (H_1): $b_i \neq 0$.

Уровень значимости 0.05

Формула t-статистики:

t - расчетное = $\frac{b_i - b_{0i}}{SE(b_i)}$

t - критическое = $t(alpha/2, n-p-1)$, так как критическая область двусторонняя.

Total slopes:

$SE(b_i)$= 2.08

t-расчетное = 1.61

t-критическое = 0.12

$(|t_{calc}| \leq t_{\text{crit}})$ => принимаем H_0.

GDP:

$SE(b_i)$ = 2.32

t-расчетное = 6.90

t-критическое = 0.00000003

$(|t_{calc}| > t_{\text{crit}})$ => отвергаем H_0 в пользу H_1.

Highest point

$SE(b_i)$ = 1.98

t-расчетное = 3.30

t-критическое = 0.002

$(|t_{calc}| > t_{\text{crit}})$ => отвергаем H_0 в пользу H_1.

Выводы значимости коэффициентов:
Коэффициент Total slopes не статистически значим
Коэффициент GDP статистически значим
Коэффициент Highest point статистически значим

Уравнение получившейся модели:

$y = 57.27 + 3.35 * x_1 + 15.98 * x_2 + 6.53 * x_3$

Коэффициент детерминации получившейся модели:

$R^2 = 0.69$

Интерпретация результатов:

При увеличении числа трасс на 1 - цена ski-pass на один день на взрослого в среднем увеличивается на 3,35 евро при всех равных прочих условиях

При увеличении ВВП страны курорта на условную единицу - цена ski-pass на один день на взрослого в среднем увеличивается на 15,98 евро при всех равных прочих условиях

При увеличении высоты пика на 1 метр - цена ski-pass на один день на взрослого в среднем увеличивается на 6,53 евро при всех равных прочих условиях

Минимальная цена ski-pass – 57,27 евро на один день на взрослого (на 2022 год это 4333 рубля)

Роли участников проекта

Горячева Екатерина - техническая часть реализации (много)
Князева Софья - план исследования и интерпретация результатов на каждом шаге (много)
Пластинина Елизавета - сбор данных и формирование отчёта (много)
Николина Варвара - сбор данных и оформление презентации (много)