МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«МИРЭА – Российский технологический университет»

РТУ МИРЭА

Институт информационных технологий (ИТ)

Кафедра прикладной математики

ОТЧЁТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 4

по дисциплине «Технологии и инструментарий анализа больших данных»

Выполнил студент группы ИКБО-20-21 Сидоров С.Д.

Проверил ассистент кафедры ПМ ИИТ Тетерин Н.Н.

Москва 2024

Практическая работа

1) Определить два вектора, представляющие собой число автомобилей, припаркованных в течении 5 рабочих дней у бизнес-центра на уличной стоянке и в подземном гараже.

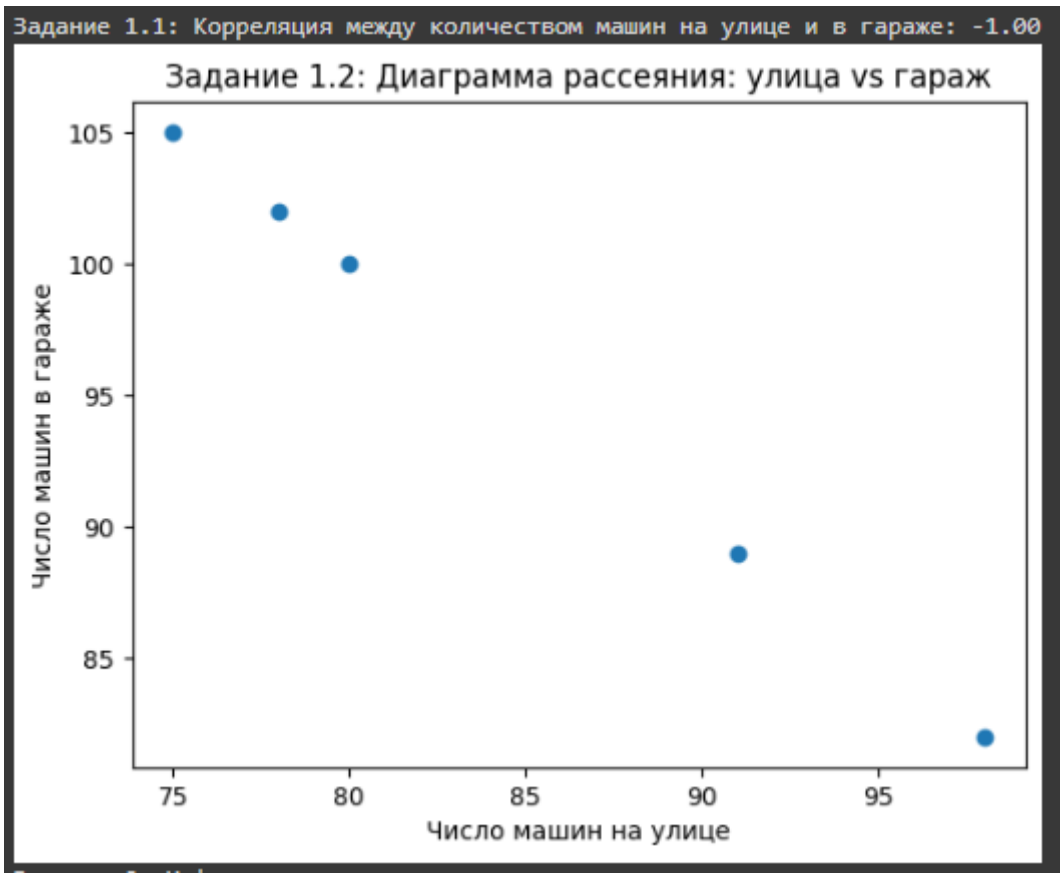
День	Улица	Гараж
Понедельник	80	100
Вторник	98	82
Среда	75	105
Четверг	91	89
Пятница	78	102

a) Найти и интерпретировать корреляцию между переменными «Улица» и «Гараж» (подсчитать корреляцию по Пирсону).

b) Построить диаграмму рассеяния для вышеупомянутых переменных

Листинг 1:

```
# Задание 1: Определить два вектора
days = ['Понедельник', 'Вторник', 'Среда', 'Четверг', 'Пятница']
street = np.array([80, 98, 75, 91, 78])
garage = np.array([100, 82, 105, 89, 102])

# Задание 1.1: Корреляция по Пирсону
correlation = np.corrcoef(street, garage)[0, 1]
print(f"Задание 1.1: Корреляция между количеством машин на улице и в гараже:
{correlation:.2f}")

# Задание 1.2: Диаграмма рассеяния
plt.scatter(street, garage)
plt.title("Задание 1.2: Диаграмма рассеяния: улица vs гараж")
plt.xlabel("Число машин на улице")
plt.ylabel("Число машин в гараже")
plt.show()
```



Вывод:

Между количеством автомобилей на улице и в гараже существует полная отрицательная корреляция (коэффициент `-1`).

Когда количество автомобилей на улице увеличивается, количество автомобилей в гараже обязательно уменьшается, и наоборот. Скорее всего, автомобилисты предпочитают парковаться либо на улице, либо в гараже, но не одновременно.

На диаграмме рассеяния точки расположены в четкой линейной зависимости, где одна переменная убывает, а другая возрастает.

2) Найти и выгрузить данные. Вывести, провести предобработку и описать признаки.

Листинг 2:

```
housing = fetch_california_housing()

df = pd.DataFrame(data=housing['data'], columns=housing['feature_names'])
```

```
df.head()
```

```
Задание 2: Информация о данных
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None
```

a) Построить корреляционную матрицу по одной целевой переменной. Определить наиболее коррелирующую переменную, продолжить с ней работу в следующем пункте.

Листинг 3:

```
# Задание 2.1: Корреляционная матрица по переменной charges
df.info()

print('\nNull values: \n', df.isnull().sum(), sep='')

corr = df.corr()['HouseAge']
corr
```

```
Задание 2.1: Корреляционная матрица
charges     1.000000
age         0.299008
bmi         0.198341
children    0.067998
Name: charges, dtype: float64
```

b) Реализовать регрессию вручную, отобразить наклон, сдвиг и MSE.

Листинг 4:

```
X = df[['HouseAge']].values
Y = df['Population'].values

def mserror(X, w1, w0, y):
    y_pred = w1 * X[:, 0] + w0
    return np.sum((y_pred - y) ** 2) / len(y_pred)
```

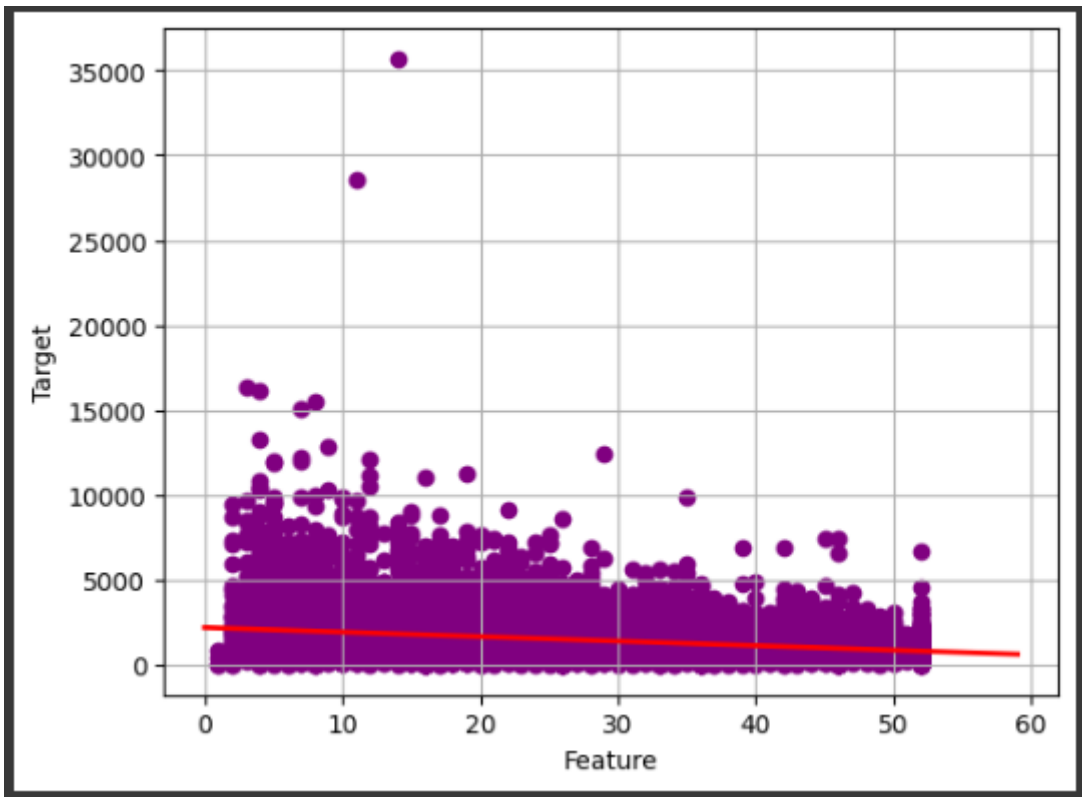
```
def gr_mserror(X, w1, w0, y):
    y_pred = w1 * X[:, 0] + w0
    error = y_pred - y
    grad_w0 = 2 * np.sum(error) / len(y)
    grad_w1 = 2 * np.sum(error * X[:, 0]) / len(y)
    return np.array([grad_w0, grad_w1])

eps = 0.0001
w1 = 0
w0 = 0
learning_rate = 0.001
n = 100000

for i in range(n):
    cur_w1 = w1
    cur_w0 = w0

    grads = gr_mserror(X, cur_w1, cur_w0, Y)
    next_w0 = cur_w0 - learning_rate * grads[0]
    next_w1 = cur_w1 - learning_rate * grads[1]

    if abs(cur_w1 - next_w1) <= eps and abs(cur_w0 - next_w0) <= eps:
        break

    w1, w0 = next_w1, next_w0

print(f"Final parameters: w1 = {w1}, w0 = {w0}")
print(f"Final MSE: {mserror(X, w1, w0, Y)}")
```

```
Final parameters: w1 = -26.647328785397143, w0 = 2188.5925608306557
Final MSE: 1169863.3501130508
```

c) Визуализировать регрессию на графике.

Листинг 5:

```
fig = plt.figure()
x = np.arange(0, 60)
our_model = w1 * x + w0
plt.plot(x, our_model, linewidth=2, color='red', label='Our Model')
plt.scatter(X, Y, label='Data', color='purple')
plt.grid()
plt.xlabel('Feature')
plt.ylabel('Target')
plt.show()
```



Вывод:

Отрицательный наклон (`-26.65`) указывает на обратную зависимость между возрастом домов и населением. Это может означать, что в более новых районах (с меньшим возрастом домов) население больше.

Большое значение сдвига (`2188.59`) показывает, что даже при нулевом возрасте домов (теоретическая ситуация) ожидается определенное базовое население.

Довольно высокое значение MSE (`1169863.35`) указывает на то, что модель имеет значительную ошибку предсказания. Это может быть связано с тем, что связь между возрастом домов и населением не является строго линейной или на нее влияют другие факторы

3) Загрузить данные: 'insurance.csv'. Вывести и провести предобработку. Вывести список уникальных регионов.

Листинг 6:

```
unique_regions = insurance_data['region'].unique()
print("Задание 3: Уникальные регионы")
print(unique_regions)
```

```
Задание 3: Уникальные регионы
['southwest' 'southeast' 'northwest' 'northeast']
```

а) Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя первый способ, через библиотеку Scipy.

Листинг 7:

```
bmi_southwest = insurance_data[insurance_data['region'] == 'southwest']['bmi']
bmi_southeast = insurance_data[insurance_data['region'] == 'southeast']['bmi']
bmi_northwest = insurance_data[insurance_data['region'] == 'northwest']['bmi']
bmi_northeast = insurance_data[insurance_data['region'] == 'northeast']['bmi']

f_stat, p_value = f_oneway(bmi_southwest, bmi_southeast, bmi_northwest, bmi_northeast)
print(f"Задание 3.1: F-статистика: {f_stat}, P-значение: {p_value}")
```

```
Задание 3.1: F-статистика: 39.49505720170283, P-значение: 1.881838913929143e-24
```

б) Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя второй способ, с помощью функции anova_lm() из библиотеки statsmodels.

Листинг 8:

```
model = ols('bmi ~ region', data=insurance_data).fit()
anova_results = anova_lm(model)
print("Задание 3.2: Результаты ANOVA")
print(anova_results)
```

```
Задание 3.2: Результаты ANOVA
                df        sum_sq      mean_sq          F        PR(>F)
region         3.0   4055.880631  1351.960210  39.495057  1.881839e-24
Residual    1334.0  45664.319755    34.231124        NaN           NaN
```

в) С помощью t критерия Стьюдента перебрать все пары. Определить поправку Бонферрони. Сделать выводы.

Листинг 9:

```
regions = ['southwest', 'southeast', 'northwest', 'northeast']
bmi_data = {region: insurance_data[insurance_data['region'] == region]['bmi'] for region in regions}
region_pairs = list(combinations(regions, 2))

p_values = []
for region1, region2 in region_pairs:
    t_stat, p_value = ttest_ind(bmi_data[region1], bmi_data[region2])
    p_values.append(p_value)
```

```
corrected_p_values = smm.multipletests(p_values, alpha=0.05, method='bonferroni')
print("Задание 3.3: Поправка Бонферрони")
print(corrected_p_values)
```

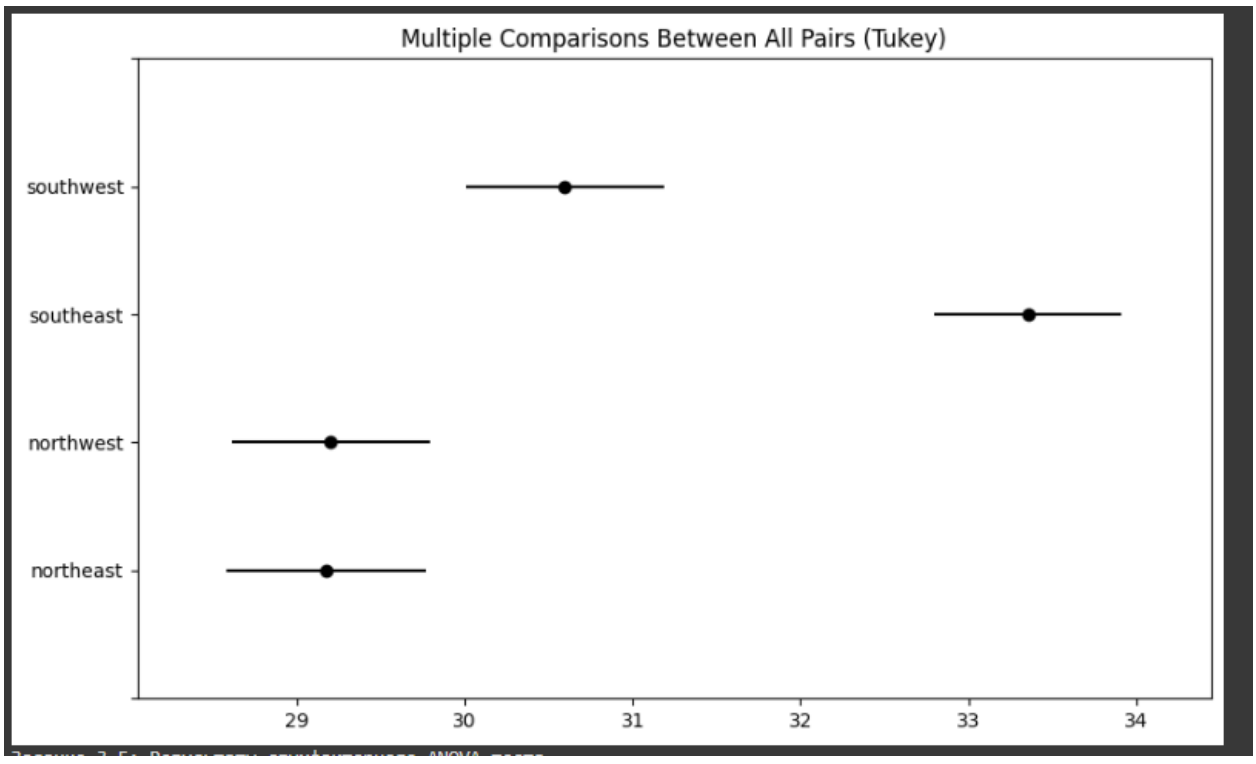
```
Задание 3.3: Поправка Бонферрони
(array([ True,  True,  True,  True,  True, False]), array([3.26244058e-08, 6.46175098e-03, 1.14516970e-02, 1.58614284e-18,
       7.11608962e-17, 1.00000000e+00]), 0.008512444610847103, 0.008333333333333333)
```

г) Выполнить пост-хок тесты Тьюки и построить график.

Листинг 10:

```
tukey = mc.pairwise_tukeyhsd(insurance_data['bmi'], insurance_data['region'])
print("Задание 3.4: Результаты пост-хок теста Тьюки")
print(tukey)

tukey.plot_simultaneous()
plt.show()
```

```
Задание 3.4: Результаты пост-хок теста Тьюки
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower   upper  reject
-----------------------------------------------------
northeast northwest   0.0263 0.9999 -1.1552  1.2078  False
northeast southeast   4.1825    0.0   3.033   5.332   True
northeast southwest   1.4231 0.0107  0.2416  2.6046   True
northwest southeast   4.1562    0.0  3.0077  5.3047   True
northwest southwest   1.3968 0.0127  0.2162  2.5774   True
southeast southwest  -2.7594    0.0 -3.9079 -1.6108   True
```



д) Выполнить двухфакторный ANOVA тест, чтобы проверить влияние региона и пола на индекс массы тела (BMI), используя функцию anova_lm() из библиотеки statsmodels.

Листинг 11:

```
model_two_way = ols('bmi ~ region + sex', data=insurance_data).fit()
anova_results_two_way = anova_lm(model_two_way)
print("Задание 3.5: Результаты двухфакторного ANOVA теста")
print(anova_results_two_way)
```

```
Задание 3.5: Результаты двухфакторного ANOVA теста
               df        sum_sq      mean_sq          F        PR(>F)
region        3.0   4055.880631  1351.960210  39.539923  1.773031e-24
sex           1.0     86.007035    86.007035   2.515393  1.129767e-01
Residual   1333.0  45578.312720    34.192283        NaN           NaN
```

е) Выполнить пост-хок тесты Тьюки и построить график.

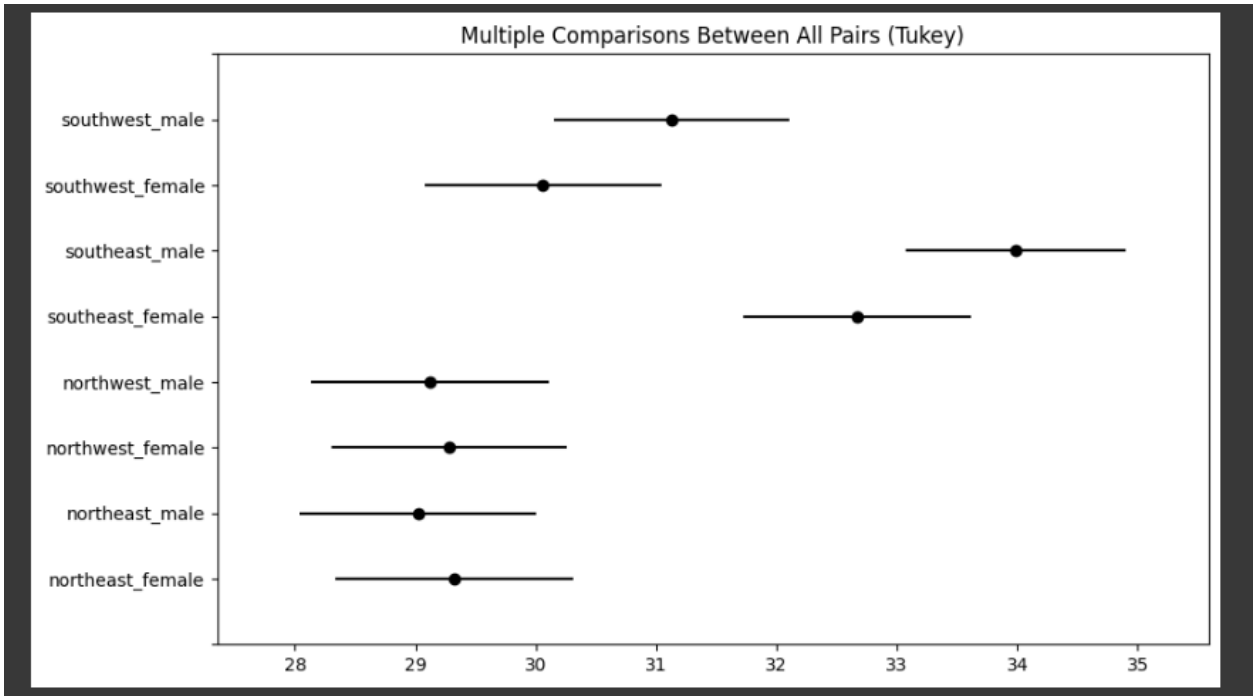
Листинг 12:

```
insurance_data['region_sex'] = insurance_data['region'] + '_' + insurance_data['sex']
tukey_two_way = mc.pairwise_tukeyhsd(insurance_data['bmi'], insurance_data['region_sex'])
print("Задание 3.6: Результаты пост-хок теста Тьюки для двухфакторного анализа")
print(tukey_two_way)

tukey_two_way.plot_simultaneous()
plt.show()
```

Задание 3.6: Результаты пост-хок теста Тьюки для двухфакторного анализа

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
northeast_female	northeast_male	-0.2998	0.9998	-2.2706	1.6711	False
northeast_female	northwest_female	-0.0464	1.0	-2.0142	1.9215	False
northeast_female	northwest_male	-0.2042	1.0	-2.1811	1.7728	False
northeast_female	southeast_female	3.3469	0.0	1.41	5.2839	True
northeast_female	southeast_male	4.6657	0.0	2.7634	6.568	True
northeast_female	southwest_female	0.7362	0.9497	-1.2377	2.71	False
northeast_female	southwest_male	1.8051	0.1007	-0.1657	3.776	False
northeast_male	northwest_female	0.2534	0.9999	-1.7083	2.2152	False
northeast_male	northwest_male	0.0956	1.0	-1.8752	2.0665	False
northeast_male	southeast_female	3.6467	0.0	1.7159	5.5775	True
northeast_male	southeast_male	4.9655	0.0	3.0695	6.8614	True
northeast_male	southwest_female	1.036	0.7515	-0.9318	3.0037	False
northeast_male	southwest_male	2.1049	0.0258	0.1402	4.0697	True
northwest_female	northwest_male	-0.1578	1.0	-2.1257	1.81	False
northwest_female	southeast_female	3.3933	0.0	1.4656	5.321	True
northwest_female	southeast_male	4.712	0.0	2.8192	6.6049	True
northwest_female	southwest_female	0.7825	0.9294	-1.1822	2.7473	False
northwest_female	southwest_male	1.8515	0.0806	-0.1103	3.8132	False
northwest_male	southeast_female	3.5511	0.0	1.6141	5.4881	True
northwest_male	southeast_male	4.8698	0.0	2.9676	6.7721	True
northwest_male	southwest_female	0.9403	0.8354	-1.0335	2.9142	False
northwest_male	southwest_male	2.0093	0.042	0.0385	3.9801	True
southeast_female	southeast_male	1.3187	0.3823	-0.542	3.1795	False
southeast_female	southwest_female	-2.6108	0.0011	-4.5446	-0.6769	True
southeast_female	southwest_male	-1.5418	0.2304	-3.4726	0.389	False
southeast_male	southwest_female	-3.9295	0.0	-5.8286	-2.0304	True
southeast_male	southwest_male	-2.8606	0.0001	-4.7565	-0.9646	True
southwest_female	southwest_male	1.069	0.7201	-0.8988	3.0367	False



Вывод:

Однофакторный ANOVA тест

Очень низкое `p`-значение (`1.88e-24 <0.05`) указывает на статистически значимую разницу в `BMI` между регионами.

Т-тест Стьюдента с поправко Бонферрони

После применения поправки Бонферрони все пары регионов, кроме `northwest-northeast`, показывают статистически значимые различия в `BMI` (`p < 0.05`).

Тест Тьюки подтверждает результаты `t`-теста с поправкой Бонферрони, показывая значимые различия между большинством пар регионов.

Пост-хок тест Тьюки

Тест Тьюки подтверждает результаты `t`-теста с поправкой Бонферрони, показывая значимые различия между большинством пар регионов.

Двухфакторный ANOVA тест

Регион оказывает значительное влияние на `BMI` (`p < 0.05`).

Пол не оказывает статистически значимого влияния на `BMI` (`p > 0.05`).

Взаимодействие между регионом и полом также не является статистически значимым (`p > 0.05`).

Существует сильная статистическая связь между регионом проживания и `BMI`.

Наибольшие различия в `BMI` наблюдаются между юго-восточным (`southeast`) и другими регионами.

Северо-западный (`northwest`) и северо-восточный (`northeast`) регионы имеют наиболее схожие показатели `BMI`.

Пол не оказывает значительного влияния на `BMI` в данном наборе данных.

Нет значимого взаимодействия между регионом и полом в отношении их влияния на `BMI`.