|  |
| --- |
| https://lh6.googleusercontent.com/QcftzNtI05T0Y6fjdSh1Rr2rt8oqZ1IvnLvbn1jLJ7CCyteVir3k-xBLv4SL1wAgWJsRhmmJSR0UW-RP63_GQenE4vVWv05BRoZTsmIcBccVTnfxwmsnNMvjg599x9SqZd8E3dkd |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **"МИРЭА - Российский технологический университет"**  **РТУ МИРЭА** |

Институт Информационных Технологий

Кафедра прикладной математики

**ПРАКТИЧЕСКАЯ РАБОТА №4**

по дисциплине

«Технологии и инструментарий анализа больших данных»

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы | ИНБО-04-20 | Ло В.Х. | |
| Проверил |  | Десятников А.А. | |
| Работа представлена к защите | «\_\_\_»\_\_\_\_\_\_\_\_2023 г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
|  |  | *(подпись студента)* |  |
| «Работа зачтена» | «\_\_\_»\_\_\_\_\_\_\_\_2023 г. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
|  |  | *(подпись руководителя)* |  |

Москва 2023

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

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание**

Эти два вектора:

улица = [80, 98, 75, 91, 78];

гараж = [100, 82, 105, 89, 102]

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

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание**

Коэффициент корреляции практически равен –1, что означает сильную отрицательную корреляцию.

1.2 Построить диаграмму рассеяния для вышеупомянутых переменных

Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описание

1. Найти и выгрузить данные. Вывести, провести предобработку и описать признаки.

Найти и выгрузить данные

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Предварительная обработка данных

Пропущенные значения отсутствуют

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Проверим пропущенные значения:

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

Описать признаки

Open: Цена открытия акции в определенный день или период времени.

* High: Самая высокая цена акции, достигнутая в тот же день или период времени.
* Low: Самая низкая цена акции, достигнутая в тот же день или период времени.
* Close: Цена закрытия акции в тот же день или период времени.
* Adj Close: Скорректированная цена закрытия. В некоторых случаях, она может отличаться от обычной цены закрытия из-за корпоративных событий, выплат дивидендов и других факторов.
* Volume: Объем торгов (количество акций), которое было куплено и продано в тот же день или период времени.

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

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

2.2. Реализовать регрессию вручную, отобразить наклон, сдвиг и MSE.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Угол наклона = 1

Коэффициент сдвига = 0.04



MSE или среднее квадратичное отклонение составило 1.03

2.3. Визуализировать регрессию на графике.

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

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

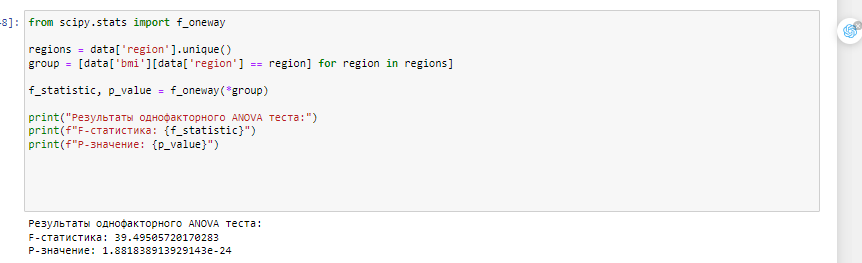
Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

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



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

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

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

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Сделать выводы

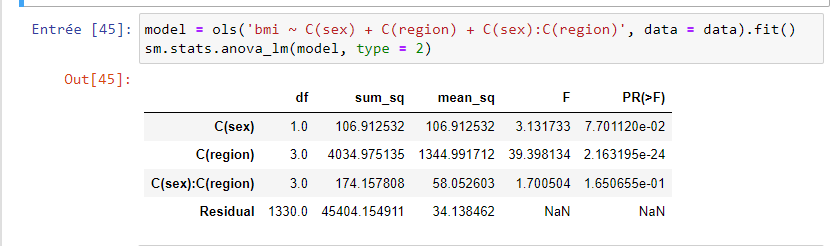
|  |  |  |
| --- | --- | --- |
| Группы | p-значение | гипотеза |
| southwest - southeast | 5.4374009639680636e-09 | отклоняется |
| southwest - northwest | 0.001076958496307695 | отклоняется |
| southwest - northeast | 0.0019086161671573072 | отклоняется |
| southeast - northwest | 2.643571405230106e-19 | отклоняется |
| southeast - northeast | 1.186014937424813e-17 | отклоняется |
| northwest - northeast | 0.951929170821864 | принимается |

3.4. Выполнить пост-хок тесты Тьюки и построить график.

Изображение выглядит как текст, снимок экрана, число, Параллельный

Автоматически созданное описание

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



3.6. Выполнить пост-хок тесты Тьюки и построить график.

Изображение выглядит как текст, Шрифт, число, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, число, документ

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

**Выводы**

После выполнения упражнения можно сделать следующие выводы:

* Корреляция помогает понять взаимосвязь между переменными.
* Линейная регрессия позволяет предсказать значения и моделировать линейную зависимость.
* ANOVA помогает проверить различия между группами или категориями. Эти методы являются важной частью анализа данных и научных исследований и широко используются в различных областях