Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

**Иркутский национальный исследовательский   
технический университет**

|  |
| --- |
| Институт информационных технологий и анализа данных |
| наименование института |

|  |
| --- |
| Отчет по дисциплине |
| «Методы анализа данных» |
| по теме: |
| «Подготовка данных для анализа» |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы |  | АСУб-20-1 |  |  |  | Устюжанин В. М. |
|  |  | Шифр группы |  | Подпись |  | И.О. Фамилия |
| Проверил преподаватель |  |  |  |  |  | Осипова Е. А. |
|  |  |  |  | Подпись |  | И.О. Фамилия |

Иркутск 2022 г.

**Содержание**

[Постановка задачи 3](#_Toc117630032)

[Выбор набора данных 4](#_Toc117630033)

[Подготовка данных для анализа 5](#_Toc117630034)

[Вывод 10](#_Toc117630035)

# Постановка задачи

1. Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта (таблица 2.1).
2. Загрузить набор данных, подготовить данные для анализа. С использованием Python выполнить корреляционный и регрессионный анализ данных. При этом необходимо:описать исходные данные;
   1. получить значения основных показателей описательной статистики и построить диаграммы рассеяния для признаков, измеренных в количественной шкале;
   2. построить корреляционную матрицу и график «тепловая карта»;
   3. проверить гипотезы о значимости коэффициентов корреляции;
   4. для заданных пар количественных признаков построить уравнения линейной регрессии, построить диаграммы остатков, оценить дисперсию остатков;
   5. для заданных пар количественных признаков найти коэффициенты уравнения полиномиальной регрессии, построить его график и построить диаграммы остатков, оценить дисперсию остатков, сравнить диаграммы остатков с диаграммами остатков для линейной регрессии;
   6. для заданных пар количественных признаков оценить индекс корреляции, сравнить со значением коэффициента корреляции;
   7. выбрать переменную отклика и переменные факторы, построить для них модель множественной регрессии, объяснить полученные значения коэффициентов множественной регрессии;
3. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

# Выбор набора данных

|  |  |  |
| --- | --- | --- |
| 11 | Forestfires.csv | Фактор – temp;  Отклик – RH |

FFMC – величина, характеризующая возможность воспламенения верхнего слоя лесной подстилки, состоящего из мелких частиц.

DMC – величина, характеризующая возможность воспламенения среднего слоя лесной подстилки (глубиной от 6 до 12 см), состоящего из лесного материала средних размеров.

DC – величина, характеризующая возможность воспламенения самых глубоких слоев лесной подстилки, состоящего из крупного лесного материала и связанной органики; является индикатором почвенных пожаров.

ISI - совмещает эффект усиления пожарной опасности ветром с оценкой пожароопасности верхнего слоя FFMC.

temp – температура.

RH – относительная влажность в процентах.

# Подготовка данных для анализа

Подключаем библиотеки, необходимые для выполнения работы.

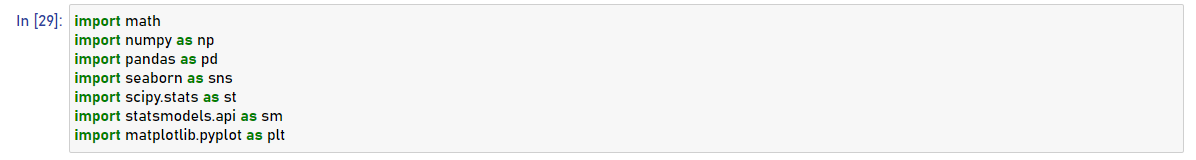


Рисунок 1 – Подключение библиотек в Python

Производим загрузку данных. Данные представляют собой 13 критериев прогнозирования лесных пожаров.

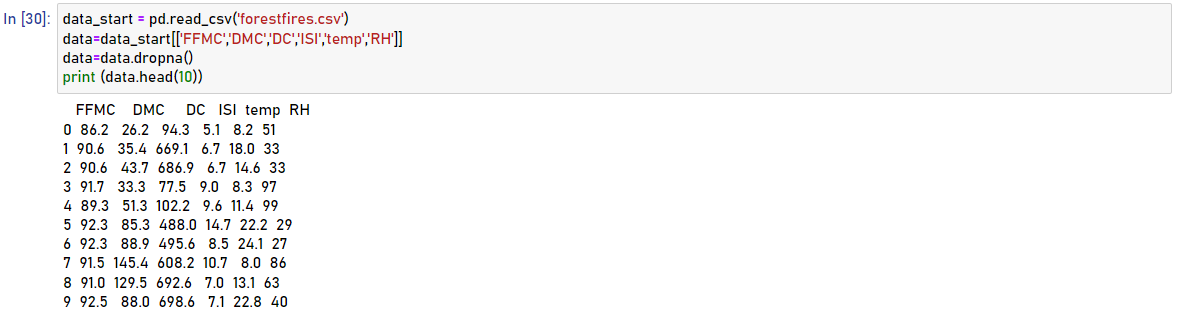


Рисунок 2 – Загрузка дата сета

Получим значения основных показателей описательной статистики и построим диаграммы рассеяния для признаков, измеренных в количественной шкале:

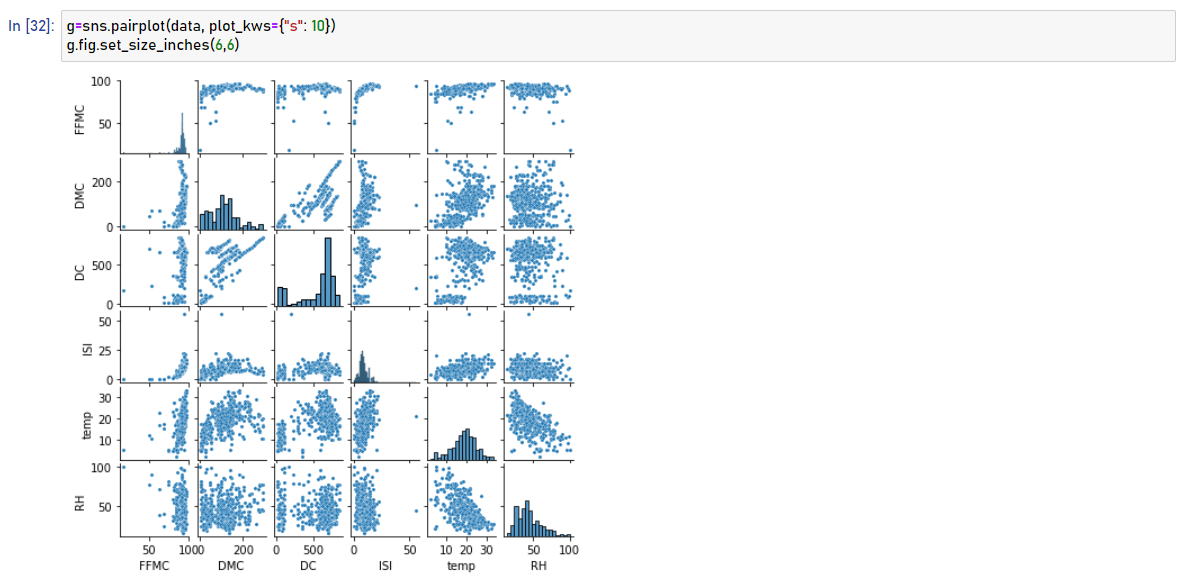


Рисунок 3 – Диаграмма рассеивания

Построим корреляционную матрицу и график «тепловая карта»:

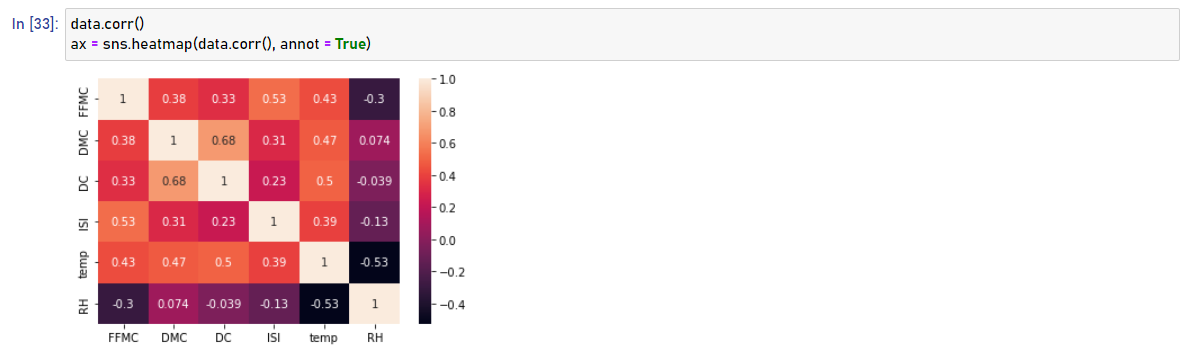


Рисунок 4 – Тепловая карта

Выдвинем нулевую гипотезу о неравенстве нулю коэффициентов корреляции тогда конкурирующая гипотеза – не равенство нулю коэффициентов корреляции.



Рисунок 5 – Значимость коэффициентов корреляции

На рисунке 5, в результате выполнения кода, получаем несколько пар – коэффициент Пирса и значение p-value которое меньше уровня значимости, а значит мы не имеем оснований принять нулевую гипотезу значит принимаем за верную конкурирующую ей.

Далее построим уравнении регрессии. Для моего варианта фактором является фактором temp, а откликом RH. Получим уравнение линейной и полиномиальной регрессии.

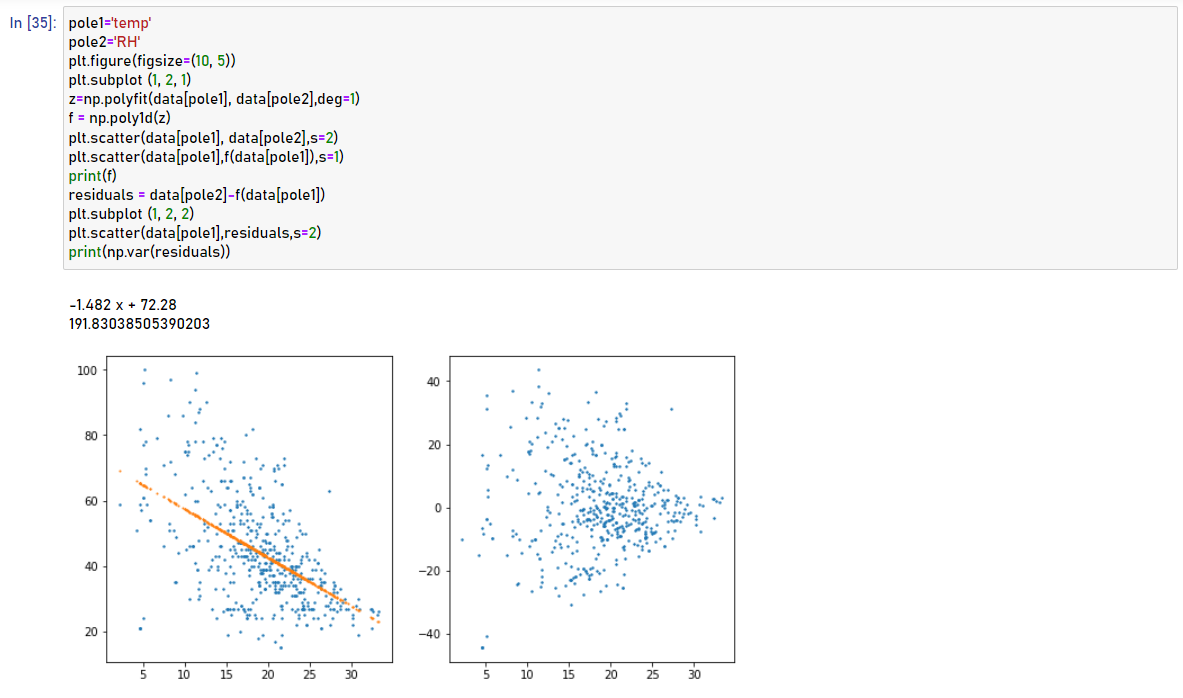


Рисунок 6 – Линейная регрессия

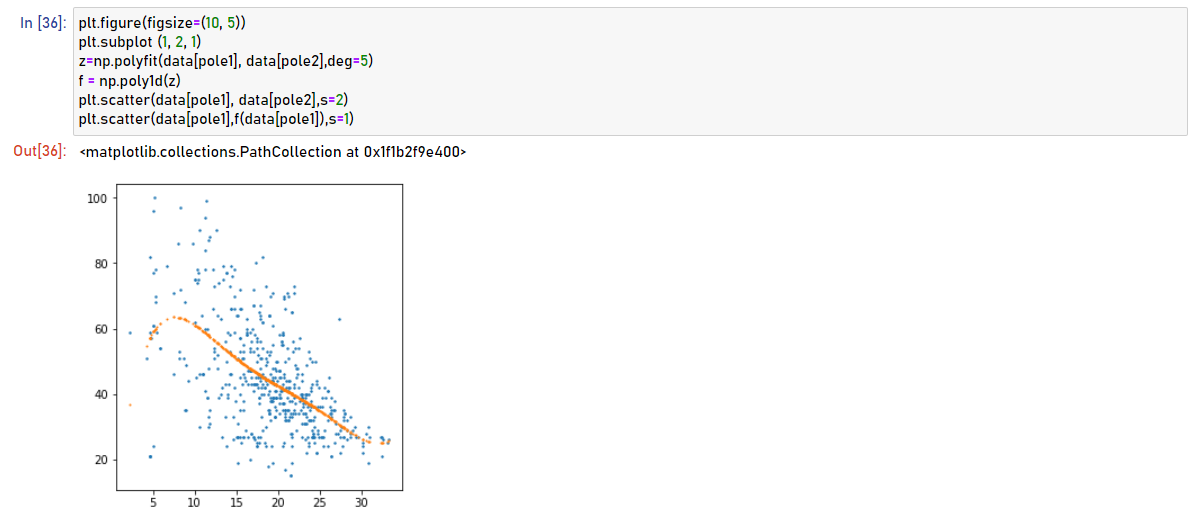


Рисунок 7 – Полиномиальная регрессия

Посчитаем индекс корреляции и коэффициент корреляции Пирсона:

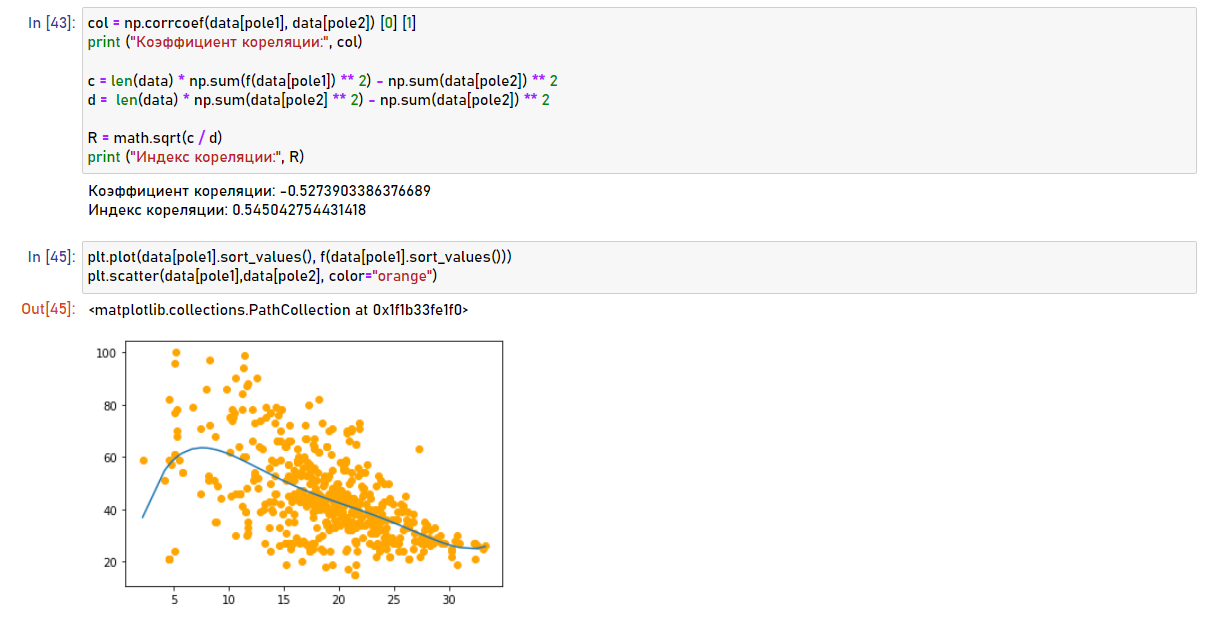


Рисунок 8 – Индекс и коэффициент корреляции

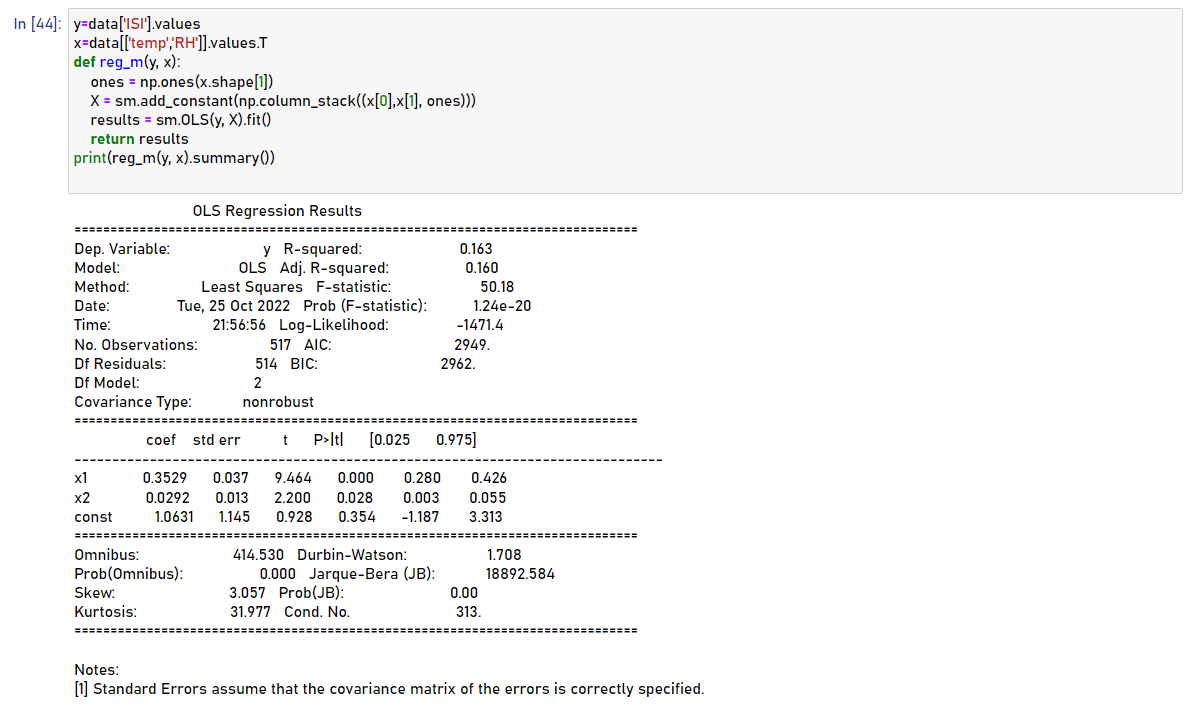


Рисунок 9 – Модель множественной регрессии

Опишем вкратце смысл значений в полученной модели множественной регрессии:

* Dep. Variable – название переменной отклика

- Model – OLS – Модель построена по методу наименьших квадратов.

- Method – Используемый метод в этой модели

- Date/Time – Дата и время создание модели

- No. Observations – Количество общих наблюдений

- Df Residuals – Количество степеней свободы

- Df Model – Количество переменных факторов.

- Covariance Type – Тип ковариации. Может быть надежной(robust) и ненадежной (nonrobust). Надежная, которая рассчитывается чтобы минимизировать или исключить переменные.

- R-squared – Коэффициент детерминации. Определяет насколько сильно переменная отклика определяется факторами.

- Adj. R-squared – Исправленный коэффициент детерминации.

- F-statistics – Критерий для проверки значимости коэффициента детерминации.

- Prob (F-Statistic) – вероятность о статистической незначимости коэффициента детерминации.

- log-likelihood, AIC/BIC – Числовые показатели для сравнения разных моделей.

- Coef – Коэффициент уравнения регрессии.

- Std err. - Оценка стандартного отклонения коэффициента.

- t statistic – критерий для оценки значимости коэффициента.

- P> | t | - вероятность о статистической незначимости коэффициента.

- [0.025 – 0.975] - являются измерениями значений наших коэффициентов в пределах 95% наших данных.

- Omnibus – мера нормальности остатков (0 – полностью нормальное)

- Prob omnibus – вероятность подтверждения гипотезы о нормальности

- Skew – мера симметрии распределения (0 – полностью симметрично)

- Kurtosis – мера остроты данных или концентрации около 0 на нормальной кривой.

- Durbin-Watson – мера равномерного распределения ошибок в наших данных

- JB – Тест Харке-Бера проверяющий нормальность.

- Cond. No - это мера чувствительности нашей модели по сравнению с размером изменений в данных

# Вывод

По результатам данной работы изучены методы корреляционного и регрессионного анализа, а также получены навыки программной реализации этих методов на языке Python.