**Министерство науки и высшего образования Российской Федерации** **Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет** **имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

**Факультет «Информатика и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

Вариант №2

Выполнил:

студент группы ИУ5-63Б

Ахтамбаев Л.Н.

Преподаватель:

Гапанюк Ю. Е.

2023 г.

Задание:

Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

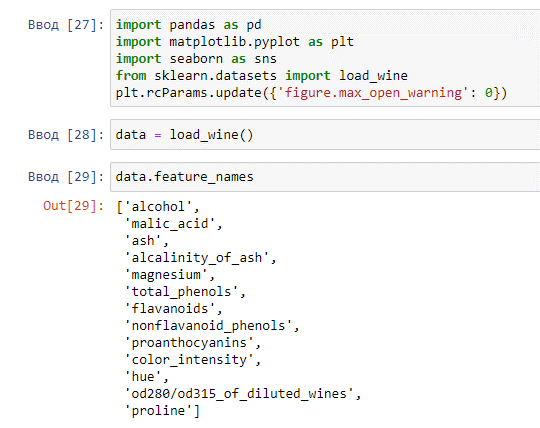
Для студентов групп ИУ5-63Б, ИУ5Ц-83Б - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Набор данных:

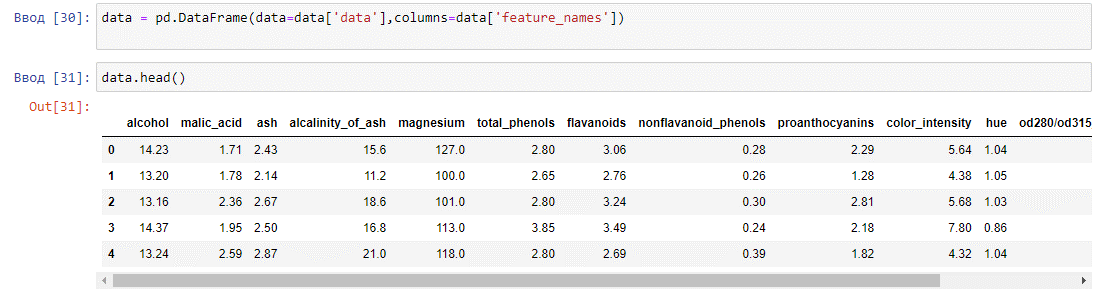
<https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris>

Решение:

Подключим все необходимые библиотеки, загрузим набор данных и проверим, что все успешно подключилось:



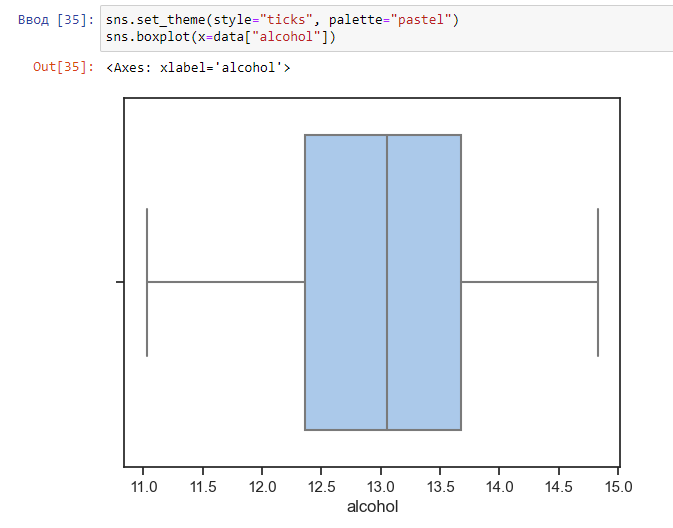
Создаем датафрейм:



Проверяем набор данных на наличие пропусков:

Пропуски отсутствуют

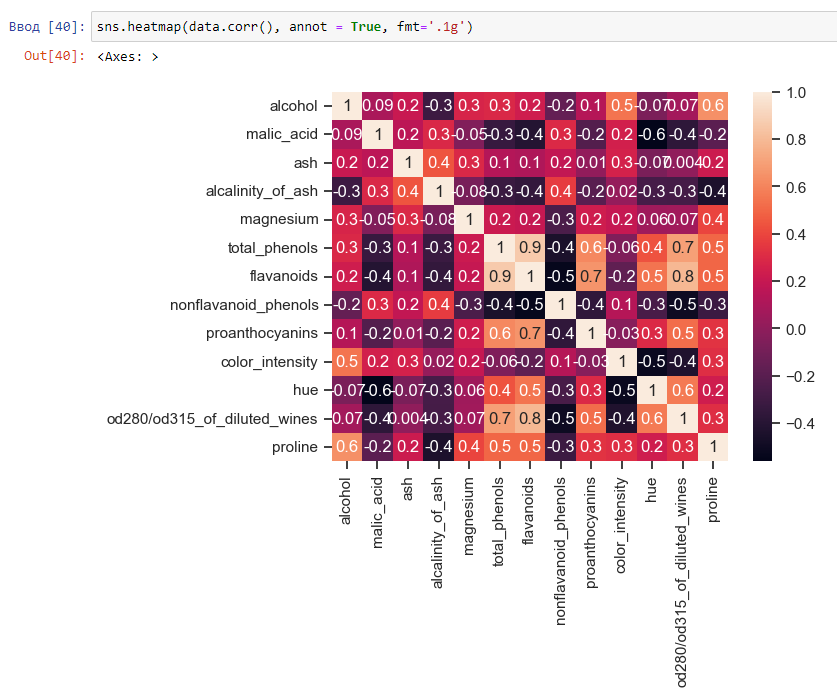
По колонке “alcohol” сделаем boxplot (Ящик с усами):



Данный график позволяет увидеть минимальные/максимальные значения, медиану, верхний и нижний квартили.

Перейдем к корреляционному анализу:

С помощью библиотеки seaborn создадим тепловую карту и изучим получившиеся значения (ниже)



По корреляционной матрице можно понять, что при построении моделей машинного обучения следует использовать признаки “flavanoids” (0.9), “od280/od315\_of\_diluted\_wines” (0.8), “total\_phenols” (0.7), “proline” (0.6) и “hue” 0.6), которые имеют наиболее сильную корреляцию с целевым признаком. Также целевой признак отчасти коррелирует с признаками “alcalinity\_of\_ash” (0.44), “proanthocyanins” (0.5), “nonflavanoid\_fenols” (0.5) и “malic\_acid” (0.44), которые мы также добавим для обучения модели. Но некоторые признаки, такие как “alcohol” (0.3), “color\_intensity” (0.3), “magnesium” (0.2) и “ash” (0.05) слабо коррелируют с целевым признаком и могут негативно сказаться на модели машинного обучения, поэтому их мы включать не будем.

Из двух сильно коррелирующих между собой признаков “flavanoids” и “total\_phenols” оставим только “flavanoids”, который имеет наибольшую корреляцию с целевым признаком.

Из вышеперечисленных признаков, которые будут включены в модель, наиболее весомый вклад окажут “flavanoids”, “od280/od315\_of\_diluted\_wines”,

“proline” и “hue”.