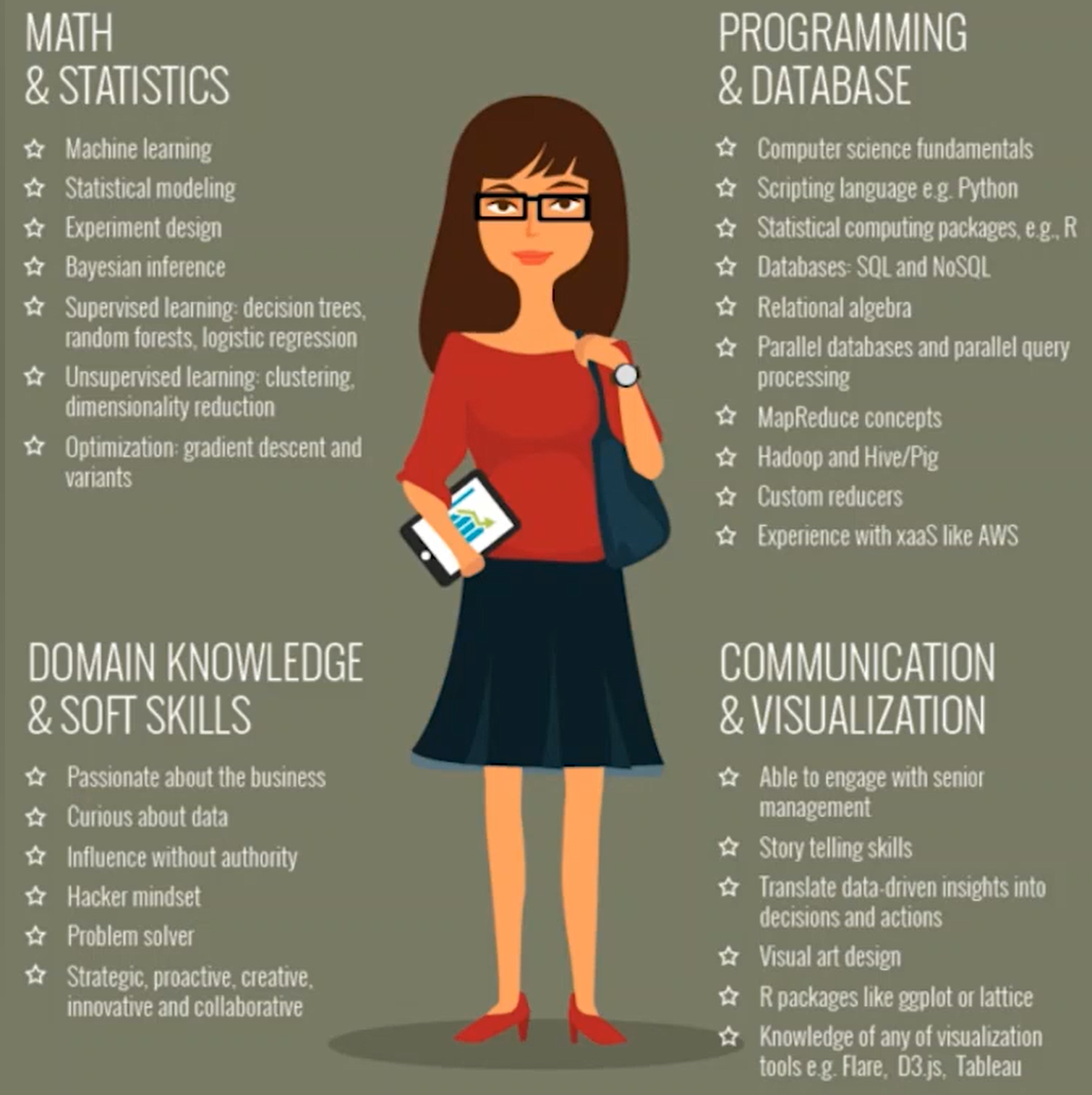


При написании данной дипломной работы я ставил перед собой следующие задачи: как устроены модели машинного обучения, как происходит процесс обучения моделей, зачем это нужно, какие ключевые концепции лежат в основании процессов, связанных с машинным обучением и какие экономические задачи можно решать, используя машинное обучение.

Методы помогают нам обрабатывать и преобразовывать данные. Атрибуты помогают нам увидеть, что хранится в данных.

DataFrame в Pandas – это не просто таблицы, но и множество атрибутов и методов к ним, состоящие из множества Series (векторов-столбцов).

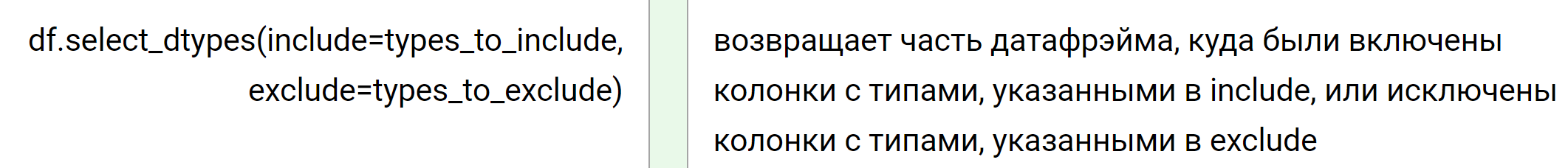


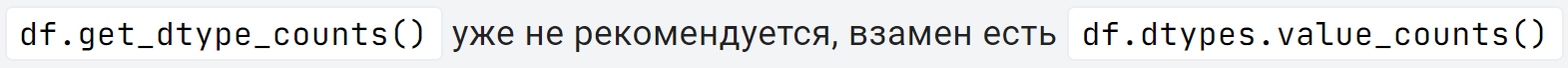
Бизнес-аналитик формируем проблемный вопрос, который нужно решить и собирает требования. Дата инженер собирает и обрабатывает данные. Дата Сайентист анализирует и подбирает модель. Разработчик машинного обучения внедряет модель в программу.

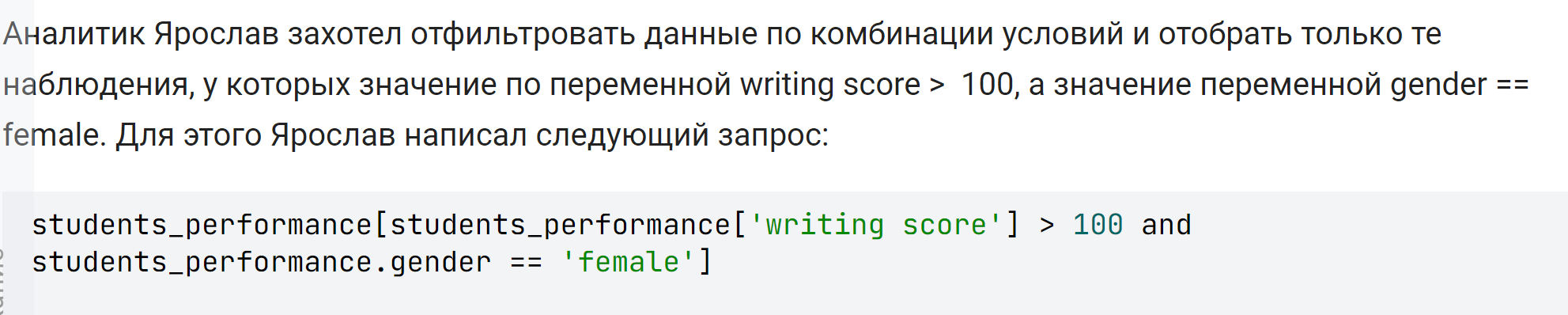
Для начала данные нужно подготовить (чистка пустых данных), формируем признаки (фичи) датафрейма, агрегируем и т.д. Далее с помощью методов машинного обучения обучаем нашу модель и прогнозируем целевую переменную. Затем мы валидируем нашу модель, проверяем, что она предсказывает верно при работе с тестовыми, а не тренировочными данными.

Подбор коэффициентов b0 и b1 – это задача машинного обучения.

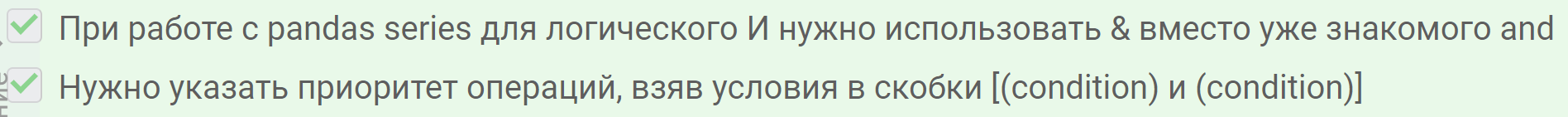
Задача всего обучения – научить модель находить общую ЗАКОНОМЕРНОСТЬ в данных, обучивших на тренировочной выборке, в которой есть некоторая частная закономерность.







<https://stackoverflow.com/questions/36921951/truth-value-of-a-series-is-ambiguous-use-a-empty-a-bool-a-item-a-any-o>



Надо просто помнить, что выражения вида `students\_performance['writing score'] > 100` выдают не логические значения, а булевы вектора; поэтому операции `and` и `or` для них не подходят. Нужна какая-то операция `and\_bool\_vector` (и `or\_bool\_vector`, соответственно), и она уже есть: `&` (и '|')

У какой доли студентов из датасэта в колонке lunch указано free/reduced?

stud\_perf['lunch'].value\_counts(normalize=True)

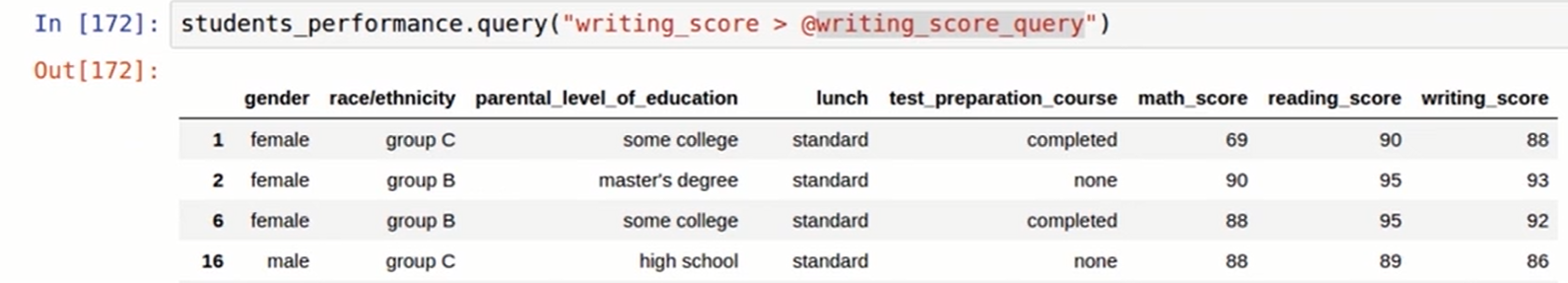
Как различается среднее и дисперсия оценок по предметам у групп студентов со стандартным или урезанным ланчем?

lunch\_std = stud\_perf[stud\_perf['lunch'] == 'standard'].describe()

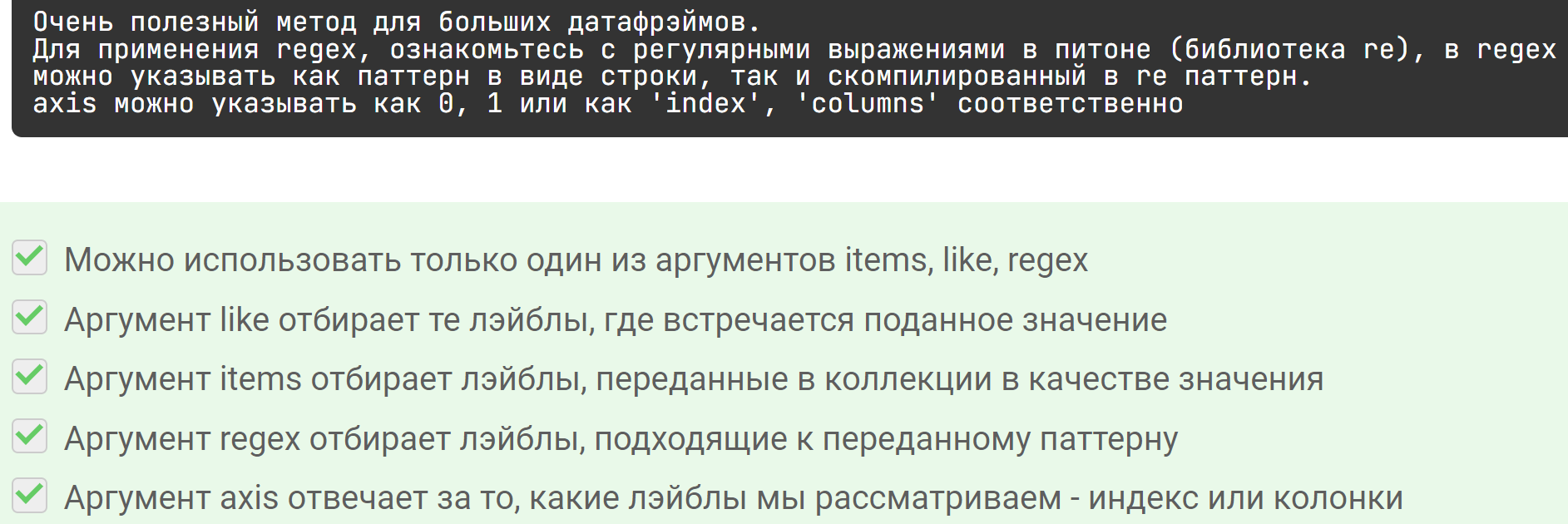
lunch\_free = stud\_perf[stud\_perf['lunch'] == 'free/reduced'].describe()

lunch\_std.compare(lunch\_free)

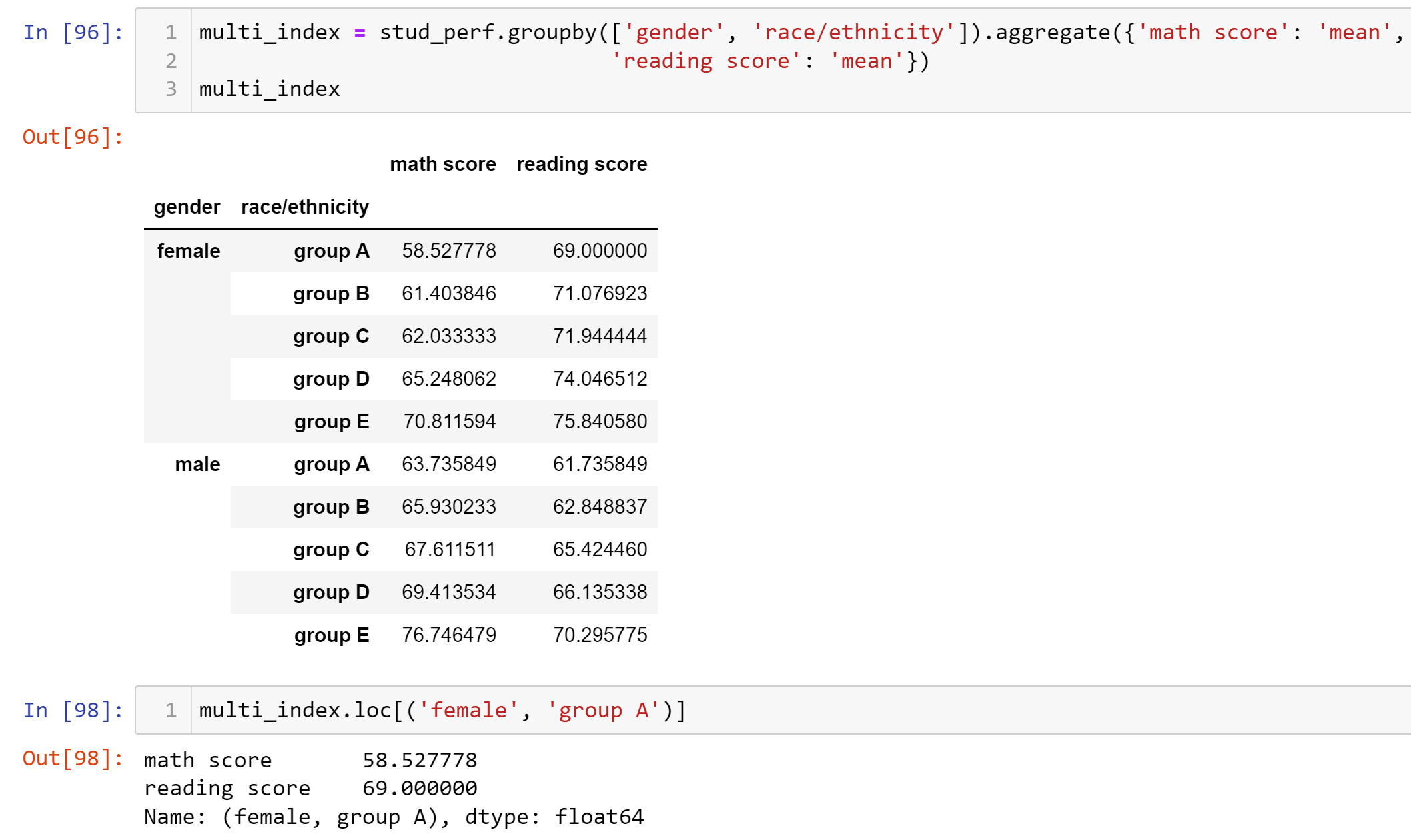
<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.compare.html>

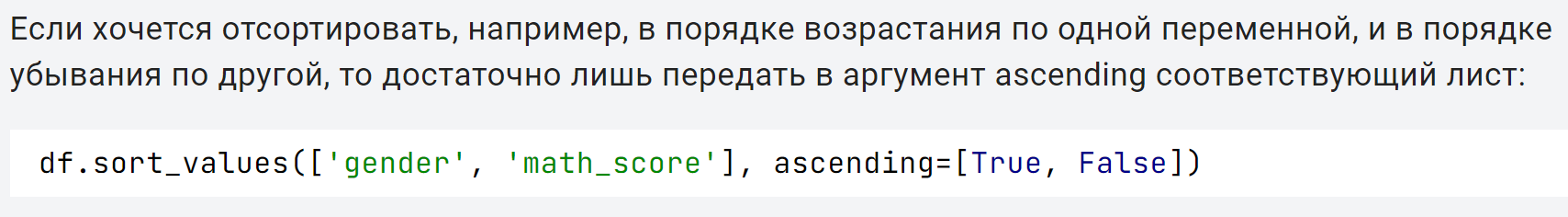


Изучите документацию метода filter и укажите правильные утверждения



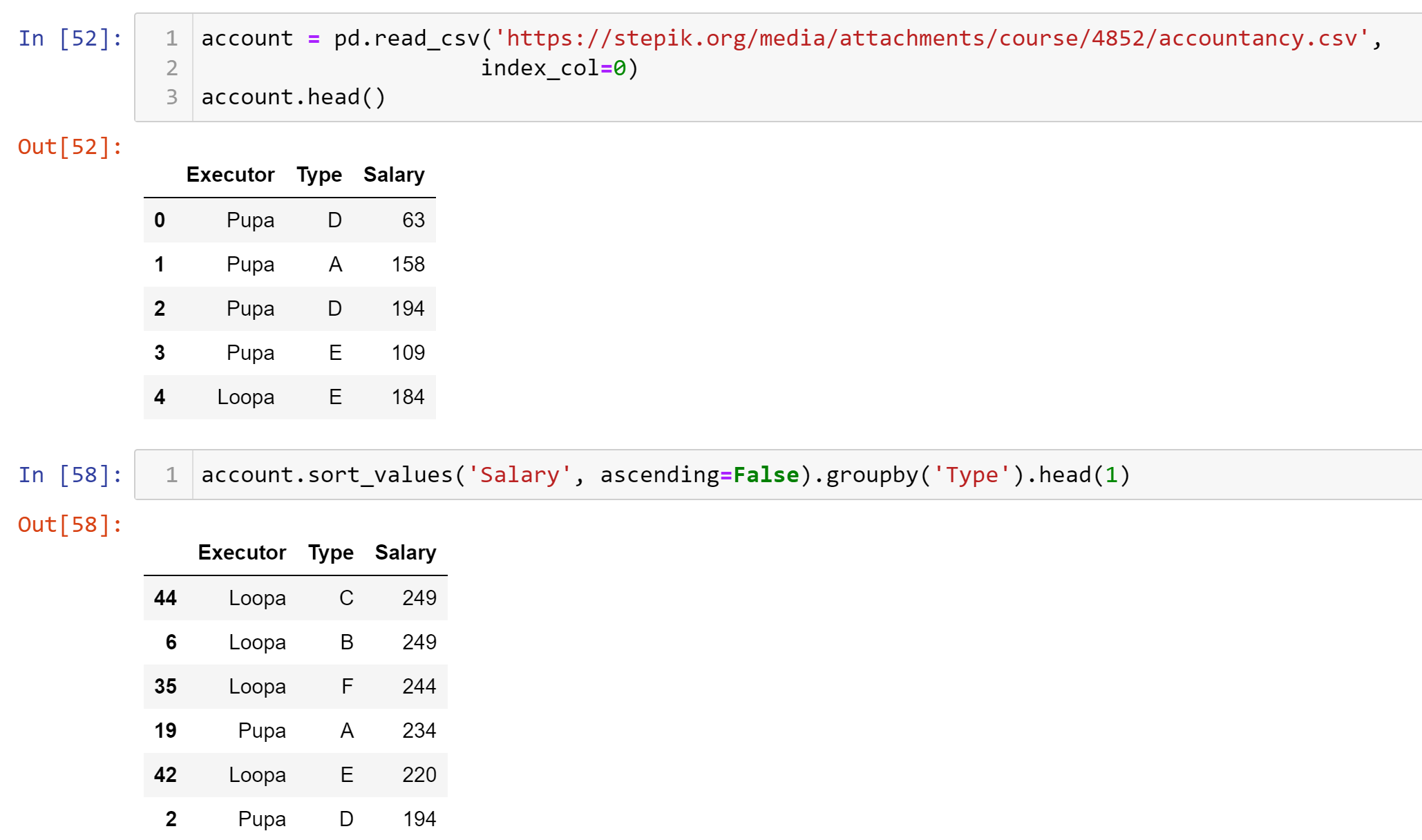
**Также небольшой совет по решению задач в целом, если вы понимаете, что не можете решить задачу, отложите ее, посмотрите несколько следующих уроков, порешайте другие задачи и потом вернитесь к проблемной задаче позже.**





К нам поступили данные из бухгалтерии о заработках Лупы и Пупы за разные задачи! Посмотрите у кого из них больше средний заработок в различных категориях (колонка Type) и заполните таблицу, указывая исполнителя с большим заработком в каждой из категорий.

<https://stepik.org/media/attachments/course/4852/accountancy.csv>



При построении карьеры в Data Science важно иметь готовые проекты.

Датасеты для анализа можно брать с Kaggle, открытых источников, или просто собирать при помощи парсеров. причём последний вариант даёт плюс к компетенциям, поскольку вы покажете умение кодить и получать данные для своих моделей.

Важно и нужно уметь отвечать на вопросы не "Сколько?", а "Почему?". Поскольку получение цифр не даёт информации, а анализ этих цифр, анализ причин почему цифры именно такие, и как они сформировались даёт более полное представление.

Работайте с API и участвуйте в хакатонах. Поднатаскайтесь в плане работы с проектами, пусть даже через API проекта. Пишите свои проекты и ссылайтесь на них в резюме.

Надо понять, как с помощью машинного обучения воздействовать на психологию пользователя и побудить его к чему-нибудь.

%matplotlib inline позволяет автоматически выводить график после его создания.

Валерий бабушкин рассказывал, что в META 10 этапов собеседований и они проходят за один день подряд. В западных компаниях очень важен софт скилл и его активно проверяют. В компаниях FAANG платят очень много (там 3 вида ЗП + помощь при релокейте и ещё при найме). 3-й вид ЗП – это выдают определённое количество акций META.

Первое место работы Анатолия Карпова было анализ данных на Stepik!) Он пытался понять причину, почему пользователи бросают курс, не проходя его до конца. Его задача была предсказать, уйдет ли пользователь с курса и как сделать так, чтобы он остался. Анатолий анализировал на каких степах у студентов возникают трудности, и они сходят с курса.

Для новичков важно даже подаваться на те вакансии, которые запрашивают более опытных сотрудников. Важно рассказать, что Вы хотите у них работать даже если у них нету Вашей вакансии, даже если Вы что-то не знаете, но готовы выучить. На собеседовании важно узнать, что ещё следует изучить и что бы посоветовали экзаменаторы.

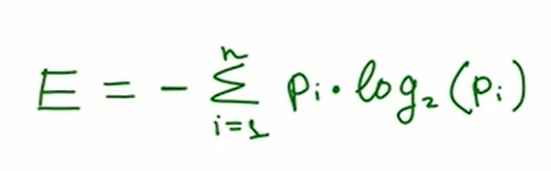
Задачу классификации также можно решить с помощью метода машинного обучения «решающие деревья».

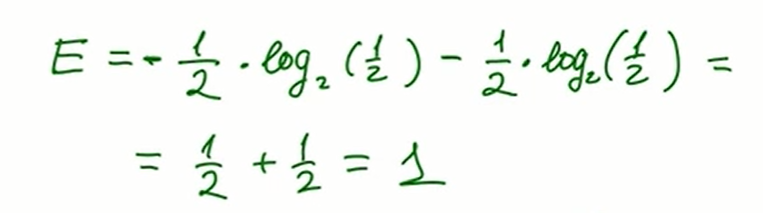
Логистическая регрессия отлично решает задачу классификацию, когда можно разделить классы при помощи прямой линии или между нашими переменными линейная взаимосвязь + нормальное распределение остатков + гомоскедастичность.

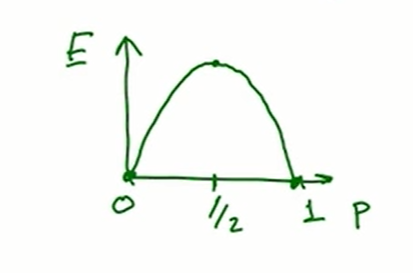
Однако все эти ограничения не имеют решающие деревья. Однако не всегда понятно какой из методов МО лучше покажет себя для решения конкретной задачи.

Каждый новый уровень (в глубине) решающего дерева снижает неопределенность (энтропию как меру хаоса), задавая правильные вопросы, которые максимально ближе подбираются к сути (условия).

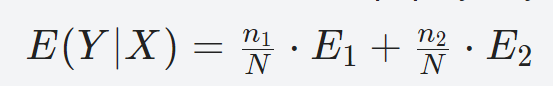
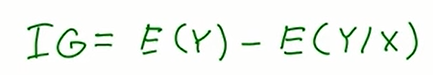
Энтропия – уровень неопределённости данных.

Формула энтропии

Пример подсчёта энтропии для двух классов с вероятностью каждого 0.5

График соотношения вероятности отнесения к классу и энтропии (при двух классах).

Выигрыш информации – разница прежней и условной вероятности.



Алгоритм ищет ту из имеющихся переменных, которая сможет снизить значение энтропии. Считаем снижение энтропии (entropy reduction) и выигрыш информации (information gain) и сохраняем данные значение и так по каждой переменной. Тогда на первый спил выберется тот, **где выигрыш информации был максимальный**. В каждой из частей мы также будем делать разделения, пока четко не сможет отнести данные к одному из двух классов (при бинарной классификации), при этом значение энтропии будет равняться нулю.