0. Обоснование избранного подхода

**Основная идея:** создание «осмысленных» (без повторов, не должны содержать только еду или только напитки) комбинаций из меню каждого из ресторанов с последующим выбором комбинаций, соответствующих определенным критериям.

**Преимущества:**

- Позволяет не просто ответить на вопрос «Какой ресторан лучше», но и предложить максимально разнообразное меню вне зависимости от выбранного ресторана.

- Критерии (как статистические, так и по отдельным показателям) можно гибко задавать, менять, быстро получать результаты, сохранять и сравнивать их.

- Данный подход может рассматриваться как задел, proof of concept для дальнейшего развития решения.

1. Логика проведения анализа

1.1. Изучение датасетов McDonalds (1) и Starbucks (2), cоответственно McD и SB

Представленность показателей, какие есть общие показатели?

**Итог:** в датасете McD намного больше показателей, ни в одном ни в другом нет данных по аллергенам (есть отдельно на сайтах).

1.2. Поиск адекватных количественных критериев для сравнения

**Итог:** найдены конкретные цифры в составе действующих в РФ нормативных документов: нормативные значения по энергетической и питательной ценности для девочек и мальчиков 14-18 лет (3), табл. 5.4. Кроме того, учтены рекомендации по распределению суточной нормы на завтраки и обеды (4), п. 6.14.

1.3. Анализ показателей

McD и SB меню и нормативные значения: какие есть пересечения, которые можно взять за основу для сравнительного исследования?

**Итог:** отобрано 6 показателей, которые есть во всех трех наборах данных:

*- Энергетическая ценность*

*- Общее содержание жиров*

*- Общее содержание углеводов*

*- Общее содержание пищевых волокон*

*- Общее содержание белков*

*- Общее содержание натрия*

1.4. Нормирование показателей.

Для каждого блюда значение показателя\* 0.3 от соответствующей дневной нормы (из предположения: на завтрак и обед должно приходиться примерно по 30% дневной нормы по энергии и питательным веществам).

**Результат:** датасет с нормированными показателями: значение каждого показателя для всех блюд выражено в доле от нормы потребления для завтрака или обеда.

1.5. Подготовка данных

5.1 Нормативы:

- выражение всех показателей по массе в единых единицах (г)

- обобщение нормы для мальчиков и девочек (показатели по пищевым волокнам и натрию для обоих полов одинаковы, остальные показатели взяты на уровне среднего значения).

**Итог:** готовы исходные данные для реализации алгоритма анализа: словарь нормативных значений 6 выбранных параметров – отдельно для завтрака и обеда (с учетом равного распределения 60% суммарной нормы: 30% на завтрак и 30% на обед).

5.2 Меню ресторанов:

- анализ категорий блюд, выделение новых блюд из более общих на основе информации о блюдах, обогащение отдельных категорий (вручную добавлены данные с сайта Starbucks – около 60 блюд в категории «Snacks & sweets»).

- создание нового признака «категория блюд» (Kind)

- очистка датасета: удаление блюд, которые содержали нулевые значения по всем 6 показателям; удаление блюд, у которых хотя бы один показатель превышал 60% от соответствующей дневной нормы (доля дневной нормы, приходящейся в сумме на завтрак обед).

- редактирование значений датасета (поиск-автозамена в редакторе Notepad++): удаление специальных знаков вроде ©, ™ и др.

**Итог:** исходные данные для алгоритма анализа: датасет меню ресторанов с 6 показателями по каждому блюду (McD и SB объединены с добавлением признака принадлежности конкретного блюда ресторану: «McD=0\_SB=1»). Датасет лучше сбалансирован по количеству позиций каждого из ресторанов.

1.6. Обзорный анализ

**Итог:** проведена оценка представленности основных категорий блюд каждого из ресторанов в датасете. На основе этих данных получены визуализации (как в абсолютом, так и в относительном выражении). Посчитаны основные статистики: mean, std, min, 25%, 50%, 75%, max по каждому из показателей (Energy, Fat, Carbohydrates, Fiber, Protein, Sodium) в разбивке по ресторанам, получены соответствующие визуализации. Результаты анализа обобщены в виде комментариев в Jupyter Notebook и описаны в разделе результатов.

2. Алгоритм анализа

2.1. Нормы: чтение в Pandas в виде словаря meal\_norm.

2.2. Загрузка словаря со взвешенной нормой (в расчете на один прием пищи), нормирование значений на основе данных словаря: показатели блюд выражены в долях от нормы в расчете на один прим пищи.

2.3. Выделение меню каждого ресторана из общего датасета. Далее для удобства работы по каждому меню применялся отдельный ноутбук с одним и тем же алгоритмом.

2.2. Разработка форматов завтраков / обедов: «1-ое блюдо, 2-ое блюда, 3-е (напиток)» и др.

2.3. Составление вручную набора групп 1-х, 2-х блюд и напитков для завтраков и обедов. Категории выбирались в соответствии с названиями («Hot Breakfast» в меню обоих ресторанов включался в состав завтраков; Lunch/Sandwiches в меню SB – в состав обедов и т.д.), а также представлениями о классическом завтраке / обеде, например, супы – на обед, салаты – как первые блюда на завтрак или обед и т.п.

2.4. Реализация алгоритма, составляющего безповторные поэлементные комбинации из элементов наборов 1-х, 2-х блюд и напитков., а также завтраков / обедов других форматов. Результат: датасеты комбинаций блюд для завтраков и обедов в 5 форматах для каждого из ресторанов – всего 10 датасетов.

2.5. Расчет суммарных значений для каждой комбинации блюд по каждому из 6 нормируемых параметров. *Например, одна из комбинаций состоит из супа, горячего сэндвича и кофе: для всех этих продуктов суммируются содержания калорийности, белков, жиров… и т.д для всех полученных комбинаций.*

2.6. Расчет средних значений и стандартных отклонений по 6 суммарным показателям каждой комбинации блюд.

2.7. Сборка всех полученных комбинаций с посчитанными метриками в единый датасет результатов.

3. Результаты

**3.1. Очистка данных**

Превышение по калориям: 1 блюдо (McD: Chicken McNuggets, 40 piece), по жирам – 3 блюда (McD: burgers), по углеводам – нет превышений, по пищевым волокнам: 2 блюда (SB, комплексные обеды), по белкам – 1 блюдо (McD: Chicken McNuggets, 40 piece), по натрию – 104 блюда (McD – 67, SB - 29). Чтобы сильно не урезать датасет, было решено повысить порог отсечения по натрию с 0.78 г до 1.0 г – количество блюд, подлежащих удалению снизилось с 104 до 73 (McD – 55, SB - 18).

**3.2. Результаты обзорного анализа предобработанного датасета**

* После очистки размер датасета снизился с 500 до 418 записей (McD: 201, из которых 63 - еда, SB – 217, из которых 136 – еда).
* **Калорийность:** группа с максимальными значениями - McD sweets/snacks (как было установлено, это серия мороженных McFlurry). Эта же группа – лидер по содержанию **Углеводов** (что ожидаемо – это почти полностью сахар).
* Другие группы с высокой калорийностью:

- SB Cold sandwiches (также лидируют по содержанию **Жиров**),

- SB Protein boxes (высокое содержание **Жиров**),

- McD Hot big breakfast.

* В среднем блюда McD имеют более высокую **Калорийность**, более высокое содержание **Жиров, Углеводов, Белков и Натрия** (McD : SB ≈ 60 : 40). Среднее содержание **Пищевых волокон** в блюдах обоих ресторанов примерно одинаковое.

**3.3. Результаты анализа сгенерированных вариантов для завтраков / обедов. Выбор лучших вариантов**

В результате применения комбинаторного алгоритма было сгенерировано более 360 тыс вариантов для завтраков / обедов, состоящих как на основе двух блюд, итак и на основе 3 блюд. Далее мы применили несколько сценариев выбора лучших блюд из полученного набора: отфильтровали варианты с определенными средними значениями (MEAN) и стандартными отклонениями (STDEV). Среднее значение = 1 соответствует норме по данному показателю в расчете на один прием пищи, среднее – 1,2 соответствует увеличению нормы на 20% и т.д.

* **MEAN = 1+/-0.1, STDEV <= 0.1** (сценарий «жесткой нормы»): Из 360 тыс вариантов только 571 прошло фильтр, из которых 298 соответствовало вариантам на основе меню McD и 223 – вариантам на основе меню SB. Таким образом, McD имеет в данном сценарии на треть больше вариантов, удовлетворяющих критерию. При учете только 3-составных завтраков и обедов у McD количество вариантов снижается до 218 против 223 SB, что дает незначительное преимущество SB. Однако, стоит отметить, что ни один из вариантов «легких» завтраков, обедов, а также перекусов на основе меню SB не прошел данный фильтр.
* **MEAN = 1+/-0.25, STDEV <= 0.2** (сценарий «мягкой нормы»). Из 360 тыс вариантов только 5538 прошло фильтр, из которых 2602 соответствовало вариантам на основе меню McD и 2936 – вариантам на основе меню SB. **В этой ситуации преимущество меню SB по разнообразию вариантов было больше уже на 12%. Аналогично,** при учете только 3-составных завтраков и обедов у McD количество вариантов снижается до 1519 McD's против 2925 SB, что дает меню SB уже почти двукратный выигрыш в разнообразии. Отметим, что рост вариантов SB обеспечивается в основном за счет группы «плотных» обедов, а также в меньшей степени – за счет «плотных» завтраков. Как и в предыдущем случае, «легкие» обеды и завтраки, как и перекусы SB не дают значимого вклада в результат.
* **MEAN = 1.2+/-0.1, STDEV <= 0.1, Energy >= 1.2** («усиленные рационы» для спортивных дней). Из 360 тыс вариантов только 1002 прошло фильтр, из которых 996 соответствовало вариантам на основе меню McD и 36 – вариантам на основе меню SB. В данном сценарии полностью доминирют варианты на основе меню McD, что ожидаемо из-за более высокой средней калорийности и содержания питательных веществ.

**3.4. Заключение. Какой ресторан выбрать?**

Ответ на данный вопрос зависит от цели и приоритетов клиента. Мы проанализировали три сценария и на основе формальных показателей можно сказать, что есликлиент ставит приоритетом сбалансированное питание, которое соответствует нормам, то разумнее будет выбрать меню McD. Если же значения нормы не столь важны (но все же учитываются) - меню SB будет гораздо предпочтительнее. Наконец, когда вам нужна высококалорийная и богатая питанием еда для обеспечения питания в спортивные дни, разумнее выбрать меню McD.

**ПРИМЕЧАНИЕ:** полученные "лучшие комбинации" на основе меню обоих ресторанов – не более чем результат работы алгоритма и фильтрация данных по формальным критериям. Мы надеемся, что этих списков достаточно, чтобы обеспечить разнообразное меню на довольно длительный срок (независимо какой ресторан клиент выберет в итоге). Большой набор данных позволяет выбрать различные варианты, полностью соответствующие запросам клиента (например, учет аллергенов в продуктах меню). Что еще можно учитывать?

* **Приоритет горячих блюд.** Например, Warm sandwiches вместо Сold sandwiches в меню SB, Hot drinks вместо Cold drinks в обоих меню.
* **Учет уровня сахара в продуктах.** Особенно это касается очень калорийных и богатых сахаром мороженных из меню McD.
* **Контроль уровня соли меню.** Ранее мы увеличили норму по содержанию натрия с 0.78 до 1.0 г из расчета на один прием пищи, поскольку это позволяло оставить больше блюд в исходном наборе. Мы считаем, что это приемлемо, т.к. в при физических нагрузках организм теряет натрий. Однако необходимо уделить особое внимание уровеню соли в продуктах, кроме того, мы рекомендуем обеспечить доступ к питьевой воде не только при приеме пищи, но и в процессе спортивных мероприятий.

**3.5. Ограничения метода**

1. Расчет на основе усредненных норм для девочек и мальчиков. Поскольку нормы потребления по энергии, жирам, углеводам и белкам для мальчиков больше (примерно на 16%), усредненные значения могут вызывать недостаточные уровни потребления для мальчиков (-8%) и избыточные для девочек (+8%). Это можно учесть и реализовать дополнительные сценарии фильтрации показателей с учетом различий норм для девочек и мальчиков. При этом, значения для пищевых волокон и натрия равны для обоих полов.
2. Как указано в задании, клиент рекомендовал использовать оригинальные наборы данных McD и SB (прилагаются к заданию)... Однако эти наборы данных основаны на меню McD и SB для США, тогда как клиент (как мы интерпретировали) находится в России или работает в российском контексте. Поэтому мы взяли значения из нормативных документов РФ.
3. Также мы обнаружили, что наборы данных меню от клиента частично устарели (это еще одна причина добавить около 60 блюд в меню SB - они уже фигурировали на сайте, но отсутствовали в данных, предоставленных клиентом).

**3.6. Предложения по дальнейшей работе**

1. Проверка других сценариев с различными фильтрами. Фильтры могут быть детализированы "по требованию" клиента.

2. Включите данные о размерах порций и потенциальных аллергенвх в продуктах в качестве новой функции и дополнительных фильтров.

3. Поиск индексов питания или расчет интегрального индекса питания. Например, оптимальное белков, жиров и углеводов = 1:1:4 или в % от калорийности: 10-15%, 30-32% и 55-60% соответственно (4). Применение подобных индексов к полученному массиву комбинаций на основе исходных меню.

4. Улучшение кода. Пересмотр кода (более "питонический") и полная автоматизация алгоритма анализа в одном ноутбуке. Обертывание всех процедур в соответствующие функции или классы.

5. Методы ML вместо ручных "сценариев": мы считаем, что методы оптимизации подойдут для

данной задачи.

6. Генератор меню с различными настройками фильтров с GUI. Можно загружать пользовательские датафреймы, получать список комбинаций и устанавливать значения критериев для выбора лучших. Если список одобрен пользователем, из отобранных в результате вариантов программа может выбрать N случайных для составления уникальных меню на много дней. Это обеспечит разнообразие блюд, которого хватит надолго.

**3.7. Использованные источники**

1. Меню ресторана [McDonalds](https://www.kaggle.com/mcdonalds/nutrition-facts)

**2.** Меню ресторана [Starbucks](https://www.kaggle.com/starbucks/starbucks-menu#starbucks-menunutrition-food.csv)

**3.** Табл. 5.4, «Методические рекомендации MP 2.3.1.2432-08: [Нормы физиологических потребностей…»](http://www.normacs.ru/Doclist/doc/11HIP.html)

4. П. 6.4, Письмо Минобрнауки России от 12.04.2012 N 06-731 «О формировании культуры здорового питания обучающихся, воспитанников» (вместе с [«Методическими рекомендациями…»](https://legalacts.ru/doc/pismo-minobrnauki-rossii-ot-12042012-n-06-731/#100010)