DAN

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ТИНЬКОФФ

ВВЕДЕНИЕ

Проект «Московское долголетие» был запущен в 2018 году. Он предоставляет пенсионерам Москвы возможность бесплатно посещать разнообразные занятия и активности в онлайн- и очном форматах. Цель проекта — помочь пожилым людям оставаться активными, сохранить свое здоровье и продлить жизнь.

В рамках проекта «Московское долголетие» москвичи старшего возраста (55+ для женщин и 60+ для мужчин) имеют возможность заниматься различными видами спорта, изучать иностранные языки, получать новые навыки и знания. Также проект помогает пожилым людям находить новых друзей, способствует общению с ними и тем самым повышает уровень социализации. Участников программы уже более полумиллиона, в вашем наборе данных представлены данные о 50 тысячах из них.

ОПИСАНИЕ ДАННЫХ

Набор данных содержит информацию о людях, вступивших в программу, и занятиях, которые они посетили в период с 01.04.2022 по 28.02.2023. За этот период состоялось более 600 тысяч групповых онлайн- и очных занятий, в датасете представлена выборка примерно из 50 тысяч участников программы, которые могли посещать эти занятия.

Одна строка в вашем наборе данных (единица наблюдения) — это описание одного участника программы. Информация о нем содержится в следующих переменных:

Уникальный номер участника	**Уникальный номер участника, целое число**
Возраст	Возраст на конец 2023 года, целое число
Пол	Пол, строка, одно из двух значений: ‘Мужчина’ или ‘Женщина’
Дата рождения	Дата рождения

Уникальный номер участника	**Уникальный номер участника, целое число**
Адрес проживания	Адрес проживания, строка
Дата регистрации	Дата вступления в программу, строка
Посещенных занятий	Количество посещенных занятий за период наблюдений, целое число
Посещенных онлайн-занятий	Количество посещенных онлайн-занятий за период наблюдений (из них), целое число
Дней посещений	Дней за время наблюдений, в которые участник посещал занятия, целое число
Число направлений	Количество посещенных направлений за период наблюдений, целое число
2022-04, 2022-05, ... 2023-02	Количество посещенных занятий за соответствующий месяц, целое число
2022-04 онлайн, 2022-05 онлайн, ...	Количество посещенных онлайн-занятий за соответствующий месяц (из них), целое число
Направление 1	Самое частое направление занятий, строка
Количество 1	Количество посещений самого частого направления занятий, целое число
Направление 2	Второе по частоте направление занятий, строка
Количество 2	Количество посещений второго по частоте направления занятий, целое число
Направление 3	Третье по частоте направление занятий, строка
Количество 3	Количество посещений третьего по частоте направления занятий, целое число

ЗАДАНИЕ

Используя имеющиеся данные, вам необходимо ответить на вопрос: «Как можно увеличить посещаемость занятий и/или привлечь больше участников в Московское долголетие?»

На основе проведенного исследования также разработайте рекомендации для создателей Московского долголетия по реорганизации занятий, расписания, привлечения потенциальных участников или любые другие рекомендации, которые помогут увеличить количество

Рекомендации по проведению исследования

Ниже приведен примерный план вашего исследования, на который вы можете опираться, и подробные комментарии к пунктам плана. Необязательно выполнять все пункты и соблюдать приведенный порядок, но работа будет оцениваться в соответствии с критериями, с которыми вам обязательно стоит ознакомиться. Постарайтесь выполнить пункты этого плана так, чтобы набрать как можно больше баллов в сумме по критериям.

1. **Предварительный анализ.** Опишите предоставленные вам данные: сколько в них наблюдений, как они распределены по разным подгруппам, какие в них есть особенности (наличие выбросов, асимметричность распределений, пропуски, неравномерность по подгруппам и тому подобное). С чем могут быть связаны эти особенности? Какие зависимости/паттерны вы наблюдаете?

2. **Постановка гипотезы.** Выявленные зависимости помогут навести вас на предположение, которое можно проверить в рамках исследования на предоставленных данных. Почему вы думаете, что ваша гипотеза может выполняться?

   Подумайте о том, что вам интересно исследовать и почему это может быть интересно кому-то кроме вас. Помните, что проверка вашей гипотезы должна быть возможна на предоставленных данных.

   Например, подумайте: как сезонность влияет на выбор активности? От каких характеристик участников программы зависит количество посещенных занятий? Как устроены предпочтения участников по направлениям занятий? Какие занятия самые популярные?

   Обратите внимание, что ответы на вопросы выше могут быть разными для разных групп пользователей. Попробуйте исследовать отдельно группы, которые можете выделить.

Организаторы Московского долголетия видят для себя приоритетным направлением именно очные занятия. Посмотрите на данные в разбивке на посетителей онлайн- и очных занятий: есть ли между ними различия? Можем ли мы выделить факторы, определяющие выбор участника? Можете ли порекомендовать что-то, что поможет организаторам перевести больше участников из онлайн-занятий в очные?

В результате этого пункта должно быть высказано предположение, которое будет проверяться в рамках исследования, — это ваша гипотеза. Кроме того, попробуйте дать конкретное обоснование выполнения выдвинутой гипотезы, почему именно зависимость работает так — это ваш механизм выполнения гипотезы. Дальше вы будете проверять выполнение гипотезы именно на предоставленном наборе данных — учтите предоставленный набор переменных при постановке гипотезы.

3. **Выбор переменных.** Выберите из представленного набора данных факторы, влияние которых вы считаете важным, и переменную, влияние на которую рассматриваете. Сформулируйте гипотезу о том, положительно или отрицательно факторы будут связаны с переменной влияния. Расскажите подробно, почему ожидаете именно такую зависимость. Обратите внимание на единицы измерения выбранных переменных при исследовании их совместного распределения.

4. **Разведывательный анализ.** Определитесь, что будет единицей наблюдения в вашем анализе. Проиллюстрируйте с помощью графиков распределение переменных, на которых вы решили сфокусироваться, если это доступно в наборе данных — посмотрите, как они меняются с течением времени. Посчитайте основные описательные статистики для них. Какие особенности вы наблюдаете? Как их можно интерпретировать, то есть объяснить? Охарактеризуйте форму распределений (скошенность, симметричность/асимметричность, вариативность).

   В ходе разведывательного анализа учитывайте наличие выбросов и пропусков в выбранных вами переменных. Стоит обратить внимание на неоднородность данных: в разных категориях распределения одной и той же переменной могут различаться. Не забывайте, что анализ можно проводить не только по исходным переменным, но и по преобразованным данным. Например, можете агрегировать участников по предпочтениям, возрасту, полу, добавить бинарных переменных — флагов, которые принимают значение 1 или 0 (например, был ли пенсионер хоть раз на онлайн-занятии или занимался ли в зимние месяцы).

   Примечание: если вам сложно сразу сформулировать исследовательский вопрос, можете начать работу над заданием с этого пункта, чтобы разобраться с предоставленным датасетом.

5. **Проверка гипотезы на данных.** Используйте математические методы и статистику, чтобы проверить выполнение гипотезы. Начните с визуального анализа и изобразите взаимосвязь на графике или нескольких графиках, если это необходимо. Учитывайте, что для разных типов переменных и взаимосвязей оптимальные визуализации различаются, и выбирайте те, что наиболее четко демонстрируют наличие или отсутствие связи.

   Проверьте, устойчивы ли ваши выводы о выполнении или невыполнении гипотезы. Посмотрите то же самое по группам или в динамике по времени — стабильно ли подтверждается результат? Если результаты получаются разными на разных подвыборках данных, предложите интерпретацию, почему так может быть.

6. **Выводы и ограничения.** Расскажите о том, как можно применить ваши результаты для пользы проекта. Какие вы можете предложить улучшающие изменения или какие получили важные знания о пенсионерах, участвующих в Московском долголетии?

   Есть ли какие-то факторы, которые вы не смогли учесть в исследовании? Могут ли быть альтернативные объяснения выполнения взаимосвязи кроме того, что вы предложили? Подумайте, какие есть ограничения в применении результатов и откуда они следуют: каких именно данных вам не хватило, чтобы быть более уверенными в результате, и как именно вы могли бы их использовать.

   Расскажите, можно ли применить полученный результат шире, не только для Московского долголетия, но и в других регионах России / странах / городах. Может быть, выводы можно расширить на людей других возрастов?

Удачи в подготовке исследования! Помните, что идеальных исследований не бывает, используйте предоставленные данные по максимуму для получения нетривиальных выводов, но не забывайте про рамки, в которых эти выводы могут быть применимы.