**Какие существуют способы соединить 2 таблицы в Excel?**

1. С помощью формул( функция ВПР, СУММЕСЛИ) (автоматически обновляется)
2. Консолидация(автоматически информация обновляться не будет)
3. Мастер сводных таблиц(правая кнопка мыши – настройка ленты – настроить панель быстрого доступа – все команды – мастер сводных таблиц) – создать таблицу на основании нескольких диапазонов консолидации – создать поля страницы – указываем те диапазоны, которые собираемся добавить в сводную - выбрать место, где будет построена таблица.

**Какие бывают типы подзапросов SQL? Приведите пример.**

Подзапрос – это обычный SQL запрос, результат которого мы можем использовать во внешнем (основном) запросе.

Существует два типа подзапросов: независимые и связанные. В независимых подзапросах вложенный запрос логически выполняется ровно один раз. Пример, :

SELECT id, name, price

FROM Products

WHERE price = (SELECT MAX(price) FROM Products)

Связанный запрос отличается от независимого тем, что его значение зависит от переменной, получаемой от внешнего запроса. Таким образом, вложенный запрос связанного подзапроса выполняется каждый раз, когда система получает новую строку от внешнего запроса. Например,:

SELECT name, price

FROM Products

WHERE 'p3' IN

(SELECT OrderNumber

FROM Orders

WHERE Orders.prodId = Products.Id);

В этом примере вложенный запрос должен логически выполниться несколько раз, поскольку он содержит столбец Id, который принадлежит таблице Products во внешнем запросе, и значение столбца Id изменяется каждый раз, когда проверяется другая строка таблицы Products во внешнем запросе.

**У Вас есть 2 таблицы: А и В. В обоих таблицах по 1 столбцу и по 10 строк. Там могут быть любые значения. Возможно, они пересекаются между таблицами, а возможно и нет. Если вы сделаете INNER JOIN этих таблиц, сколько может получиться минимум и максимум строк в результирующей таблице? А если LEFT JOIN? Объясните, почему получается такой результат.**

Если мы сделаем INNER JOIN этих таблиц, то минимальное количество строк в таблице будет 0, если нет пересекающихся значений между таблицами вообще, а максимальное значение 10, если все значения полностью одинаковы для двух таблиц.

Eсли LEFT\_JOIN (SELECT A.id FROM A LEFT JOIN B ON A.id = B.id;), то для данных таблиц мы всегда будем в результате видеть значения из таблицы A. Это происходит, так как  условие  LEFT JOIN  позволяет выбирать строки обеих таблиц, которые совпадают, плюс все строки из левой таблицы (A) даже без совпадения со строками правой таблицы (B).

**Напишите регулярное выражение, которое будет соответствовать любому российскому мобильному телефону в любом из следующих форматов записи:**

**+70000000000**

**+7 000 000 00 00**

**+7 (000) 000 00 00**

**80000000000**

**8 000 000 00 00**

**8 (000) 000 00 00**

^(8|\+7)([\ ]?)(\(?\d{3}\)?[\ ]?)?[\d\ ]{7,9}$

**Напишите SQL запрос к таблице users, который вернёт список из 100 наиболее часто встречающихся имён, написанных кириллицей. Таблица users содержит столбцы: id, email, name**

SELECT name

FROM users

WHERE name REGEXP '[Α-Яа-я]'

GROUP BY name

ORDER BY count(name) DESC

LIMIT 100;

**Объясните простыми словами что такое лямбда-функция в Python. Приведите пример кода, который с помощью лямбда-функции добавит к pandas-датафрейму новый столбец, оперируя данными из уже имеющихся столбцов.**

Лямбда-функции – функции, которые не имеют имени, то есть являются анонимными. Как известно, ключевое слов def используется в Python для определения обычной функции. В свою очередь, ключевое слово  lambda  используется для определения анонимной функции. Основная причина применения **лямбда** **функций** — создание функции, которая используется в коде единожды. Использование таких функций позволяет снизить число строк кода, которые вам придется написать. Эта функция может иметь любое количество аргументов, но вычисляет и возвращает только одно значение

Пример создания столбца стоимости жилья за кв.метр ‘price\_per\_sq’ по имеющимся признакам цены и площади жилья:

data['price\_per\_sq'] = data.apply(lambda row: row["price"] / row["sqft\_lot"], axis=1)

**Как и зачем используется %s в Python? Приведите пример.**

%s в Python используется для включения данной строки в другую строку. Этот оператор автоматически обеспечивает преобразование типа заданного значения в строковый тип данных.

str1 = "Юлия"

str2 = "аналитиком"

str3 = "UCHI.RU"

print("Здравствуйте! Меня зовут %s! Я хотела бы работать %s в компании %s." %(str1, str2, str3))

**Что не так с этим кодом на Python?**

**m = 0**

**def pp():**

**m += 1**

**pp()**

Python ожидает, что m внутри функции будет локальной переменной. Соответственно, он ищет, где она объявляется в функции. А она не объявлена. Python падает с ошибкой. Также функция ничего не возвращает. Предлагаю добавить аргумент m в функцию pp, вернуть значение m и вывести на экран результат работы функции:

def pp(m):

m += 1

return m

m = 0

print(pp(m))

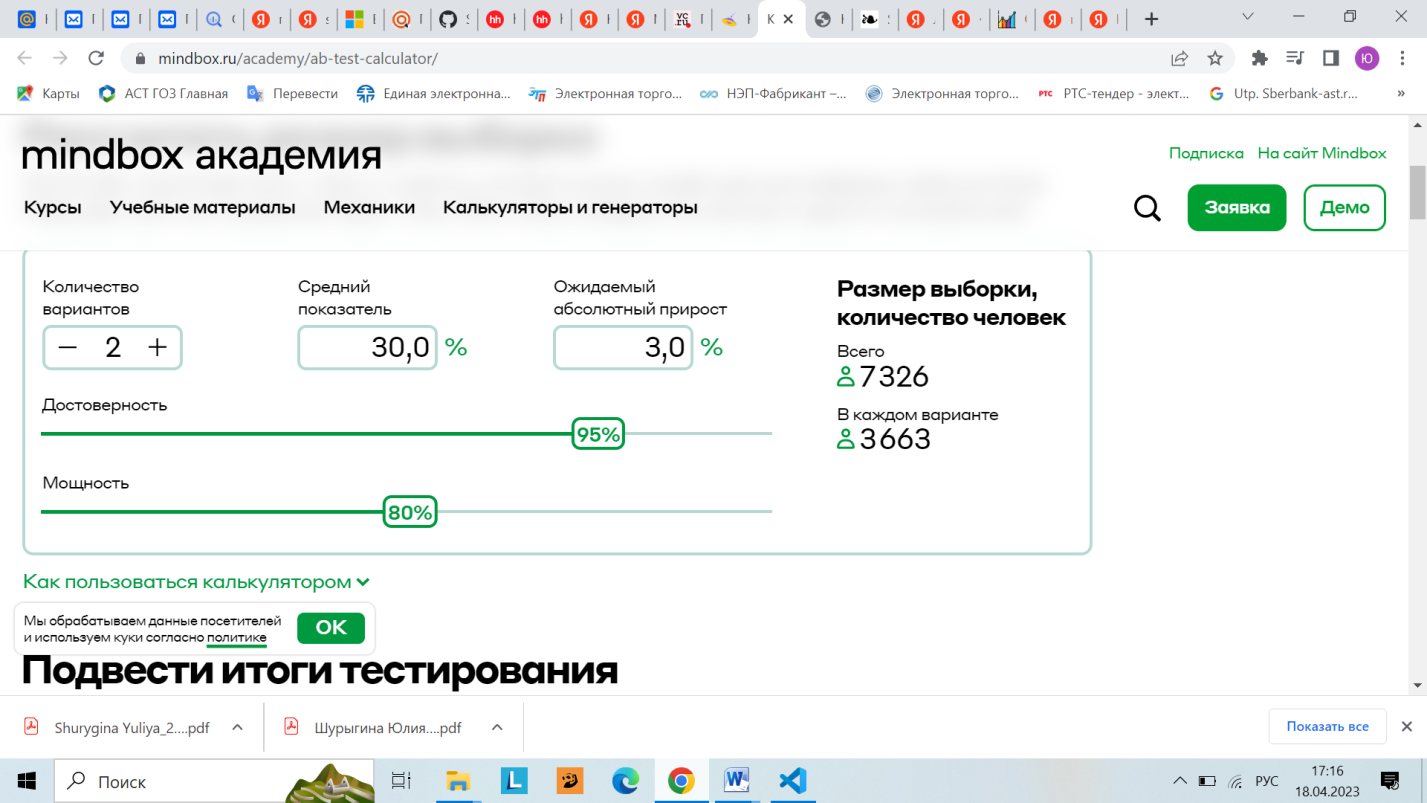
**Объясните кратко и простыми словами, что такое статистическая мощность и уровень значимости.**

Уровень статистической значимости α – вероятность ошибки первого рода. Чаще всего ее принимают равной 0,05. Это означает, что мы в 5 % случаев разрешаем ошибиться, а именно принять альтернативную гипотезу, когда на самом деле верна нулевая. 5 % вероятность, что мы найдем отличия, когда их на самом деле нет.

Статистическая мощность (1- β) – это вероятность найти отличия, когда они на самом деле есть. Мощность теста должна быть не менее 80 %.

**Мы хотим протестировать изменение на странице, которое, как мы ожидаем повлияет на среднее время просмотра этой страницы. Сколько нужно посетителей в выборках, чтобы результат A/B теста был достоверен?**

Чтобы рассчитать размер выборки, с которой вы получите статистически значимые результаты, нужно знать 4 вещи:

* **среднее время просмотра***, мин*
* *статистическая мощность — 80%;*
* *уровень  статистической значимости — 5%,*
* *****ожидаемый прирост среднего времени просмотра .*

Например, примем среднее время просмотра 30 мин, ожидаемый прирост 3 %. Тогда с помощью калькулятора( например mindbox), вычисляем размер выборки 7326 посетителей.(в каждой группе 3663 чел)

**Что такое выбросы и как с ними обращаться?**

**Выбросы - данные, значительно отличающиеся от остальной части набора данных.** Другими словами, выброс содержит значение, которое несовместимо или не соответствует общему поведению.

В начале исследования данных необходимо обнаружить и обработать выбросы.

Мы можем выбирать из множества подходов, в зависимости от имеющегося набора данных. Одним из самых простых и наглядных методов обнаружения выбросов является визуализация данных( например, ящик с усами).

Далее смотрим, как распределены данные в нашей предметной области и в зависимости от этого выбираем метод фильтрации данных. Иногда имеет смысл перепробовать несколько способов замены выбросов и посмотреть, как ведут себя несколько разных моделей.

**Какие методы можно использовать для заполнения пропущенных данных в датасете?**

**Одномерные методы** (Single Imputation) — это заполнение с использованием данных одного столбца. Другими словами, чтобы заполнить пропуски мы берем данные того же признака.

**Количественные данные**.

Самый простой способ работы с пропусками в количественных данных — заполнить пропуски константой. Например, нулем (подходит для алгоритмов, чувствительных к масштабу признаков).

Также количественные данные можно заполнить средним арифметическим или медианой (Statistical Imputation), причем с предварительной группировкой данных можем получить более близкие значения.

**Категориальные данные**. Для категориальных признаков в некоторых случаях можно провести дополнительное исследование, а также применить **метод заполнения наиболее часто встречающимся значением** (модой). Если пропусков немного, этот метод вполне обоснован. При большом количестве пропусков, можно попробовать создать на их основе новую категорию.

Более продвинутый подход — **многомерные методы** (Multivariate Imputation), заполнение пропусков одной переменной на основе данных других признаков. Другими словами, мы строим модель машинного обучения для заполнения пропусков.

Сэмплирование представляет собой такой процесс обработки данных, при котором общее представление об информации выстраивается на основе некоторой ее части — из небольшой выборки.

Существует два типа методов выборки:

Вероятностная выборка предполагает случайный отбор, что позволяет делать убедительные статистические выводы о всей группе.

Существует четыре основных типа вероятностной выборки.

1. Простая случайная выборка

В простой случайной выборке каждый член совокупности имеет равные шансы быть отобранным. Можем использовать такие инструменты, как генераторы случайных чисел или другие методы, полностью основанные на случайности.

1. Систематическая выборка

Систематическая выборка похожа на простую случайную выборку, но обычно ее немного легче проводить. Вместо того, чтобы генерировать случайным образом числа, люди выбираются через равные промежутки времени.

сотрудников.

1. Стратифицированная выборка

Стратифицированная выборка включает в себя разделение совокупности на подгруппы, которые могут существенно различаться. Затем вы используете случайную или систематическую выборку, чтобы выбрать образец из каждой подгруппы.

1. Кластерная выборка

Кластерная выборка также предполагает разделение совокупности на подгруппы, но каждая подгруппа должна иметь сходные характеристики со всей выборкой. Вместо выборки отдельных лиц из каждой подгруппы вы случайным образом выбираете целые подгруппы.

Невероятностная выборка предполагает неслучайный отбор на основе удобства или других критериев, что позволяет легко собирать данные.

**4 вида невероятностной выборки:**

1. Удобная выборка

Удобная выборка просто включает людей, которые оказались наиболее доступными для исследователя.

1. Выборка добровольных ответов

Вместо того, чтобы исследователь выбирал участников и напрямую связывался с ними, люди добровольно выступают сами (например, отвечая на общедоступный онлайн-опрос).

1. Целенаправленная выборка

Этот тип выборки, также известный как выборка суждения, предполагает, что исследователь использует свой опыт для выбора выборки, которая наиболее полезна для целей исследования.

1. Выборка методом снежного кома

Если доступ к популяции затруднен, можно использовать выборку методом снежного кома для набора участников через других участников.

**Какой областью в PA/DE/DS/ML Вы хотели бы заниматься? Почему?**

**На данный момент вижу себя продуктовым или дата аналитиком, так как мне интересно** анализировать пользовательские метрики, понимать, как покупатели взаимодействуют с продуктом, выяснять, что заходит, а что нет, проводить A/B тесты и оценивать результаты, то есть активно **участвовать в развитии продукта. Data Science является для меня следующей ступенью развития, когда захочется углубиться в построение моделей и математику.**

**Как Вы думаете, что характеризует хорошего data-специалиста?**

# Критическое мышление, аналитический склад ума, ответственность, скрупулёзность, системность.