Какие предположения относительно данных и их анализа Вы использовали при написании программы?

1. Анализируя предоставленные данные, я пришел к выводу, что они содержат информацию о клиническом исследовании. Испытуемые разделены на две группы, Treatment group 1 и Treatment Group 2. Судя по всему, одна из этих групп является контрольной группой, то есть исследование плацебо-контролируемое.
2. Данные содержат измерения по двум параметрам – EFF01 и EFF02 (параметры эффективности). Исходя из данных, непонятно, что измеряют данные параметры, но понятно, что они используются для статистического сравнения влияния какого-то воздействия (судя по всему препарата) на испытуемых для двух групп. Так как задача состоит в составлении аналитики по Parameter 1, мною было решено не использовать Parameter 2 для дальнейшего анализа данных.
3. Данные содержат два флага - Intention to treat population flag и Per protocol population flag, которые говорят о том, входит ли испытуемый в соответствующую популяцию (ITT-популяция и PP-популяция). Значение 1 – входит, значение 0 – не входит.

Насколько я понял, ITT-популяция клинических испытаний предназначена для отражения того, что можно было бы увидеть, если бы лечение использовалось в клинической практике. PP-популяция обычно определяется как все пациенты, завершившие исследование без серьезных отклонений от протокола, то есть те, кто следовал правилам исследования.

Несмотря на то, что необходимо привести аналитику по ITT-популяции, мое первое предположение заключалось в том, чтобы исключить из анализа тех испытуемых, которые имели серьезные отклонения от протокола, то есть не входящие в PP-популяцию.

Немного углубившись в данную тему, я понял, что данное решение не соответствует дизайну клинических испытаний. Суть ITT-популяции в исследовании как раз заключается в том, чтобы отразить реальную эффективность, а так как в реальной жизни не все люди принимают препараты по всем предписаниям и могут прерывать лечении, исключать испытуемых из анализа, не входящих в PP-популяцию нецелесообразно.

В связи с этим, для составления отчета я использовал данные обо всех пациентах, входящих в ITT-популяцию.

4. Данные содержат значения NaN в столбце «AVAL». Мое предположения заключается в том, что это так называемая «Missing Data in Clinical Trials». В руководстве «International Conference on Harmonisation (ICH) E9 guideline (1998)» упоминается предотвращение пропуска данных; но также признается, что не существует единого способа обработки недостающих данных из-за уникального дизайна и характеристик измерения. В связи с этим, я решил использовать один из популярных подходов для решения данной проблемы. А именно – использование репрезентативных значений для замены пропущенных значений. Выбор пал на среднее арифметическое, которые вычисляется по всем значениям той же Treatment group и соответствующему визиту. Такое решение было принято на основании того факта, что нужно предоставить статистику не по каждому отдельному испытуемому, а по всем значениям для конкретного визита и конкретной группы. Более того, среднее, вычисленное по такому правилу, будет все более репрезентативно по мере увеличения количества испытуемых.
5. Результаты по параметру 1 для каждого испытуемого есть для 1, 2, 3 визита. Исключением является испытуемый под номером 313, для которого данные ограничиваются первым визитом. Предположение заключается в том, что если бы данные о 2 и 3 визите для данного испытуемого были упущены, то они бы были помечены как NaN. Но так как данные просто отсутствуют, и данный пациент входит в PP-популяцию, было принято решение не задавать значения для пропущенных визитов путем вставки репрезентативных значений.

Как Вы думаете, можно ли было бы выполнить вычисления из тестового задание 1 в Excel? Какие были бы достоинства и недостатки подобного решения?

Я не имею большого опыта работы с Excel для анализа данных, поэтому не смогу развернуто ответить на данный вопрос. Несмотря на это, учитывая те особенности, с которыми я знаком, могу сделать некоторые предположения.

Преимущества:

- Наличие необходимых встроенных статистических функций, не требующих отдельной установки. Данная особенность позволит быстро выполнить необходимые вычисления из тестового задания.
- Не нужно писать программный код. Данных факт дает возможность быстро сделать необходимые вычисления. На мой взгляд это проще, чем делать те же вычисления, используя, например, Python.

Возможные недостатки:

- Очень сильно влияет «человеческий фактор» на корректность вычислений. На мой взгляд в Excel намного меньше способов проверить правильность вычислений, чем при работе на Python.

Резюмируя, я считаю, что выполнить необходимые вычисления в Excel можно. Это будет быстрее, чем, например, на Python, но нужно быть очень внимательным, чтобы не допустить ошибки.

Как Вы считаете, в каких случаях программист может сам принять решение о выборе подхода для анализа данных, а когда обязательно нужно запросить подтверждение?

На мой взгляд, программист должен запрашивать подтверждения, когда ему не хватает знаний в предметной области и когда он не уверен можно ли применять те или иные операции к данным. Рассмотрим, например, тестовое задание. В данных существуют значения NaN, мною было выбрано решение заменить эти данные на среднее. Но если смотреть в общем, цель аналитики по клиническим исследованиям – предоставить наиболее приближенную к реальности модель воздействия препарата на людей. Так как дизайн каждого отдельного исследования индивидуален и запротоколирован, программисту стоит следовать предписанным исследователями правилам. В связи с этим, ему стоит запросить подтверждение, чтобы узнать, как в данном конкретном случае обрабатывать пропущенные значения. То же самое касается пропущенных визитов. Как конкретно обрабатывать эти пропуски, на мой взгляд, должно быть прописано.

Если постараться сказать более обобщенно, то программисту следует запрашивать подтверждение, если конечный аналитический отчет подразумевает конкретное решение. То есть, существуют четкие однозначные предписания, следуя которым, должны получиться единственно верные результаты. Напротив, если аналитическая задача подразумевает более «свободные» результаты, то программист должен принимать решения сам. Таким образом он будет привносить что-то новое, до чего остальные аналитики могли не догадаться.

По Вашему мнению, какие подходы к контролю и обеспечению качества программирования могут быть эффективны для задач анализа данных клинических исследований?

Во-первых, это вышеописанное подтверждение своих действий. Например, обязать программиста подробно расписывать свои действия для решения той или иной аналитической задачи. Если описанные им шаги в чем-либо не соответствуют логике исследования, то необходимо их скорректировать. Это позволит избежать тех ошибок, которые потом будет тяжело исправить.

Также, программисту необходимо следить за тем, какие ошибки он допускает в решении задач и анализировать их. Можно, например, вести список своих ошибок с подробным описанием, это позволит не допускать такие же ошибки в будущем.

Другим подходом является групповое решение поставленных задач. Если одну аналитическую задачу будет решать группа людей, то программисты смогут корректировать решения друг друга и дополнять их.

Также нужно учитывать опыт предшествующих задач. Это позволит не допускать тех ошибок, которые уже были допущены, и перенять опыт в решении.