Проектное задание №1

Итак, ваше первое проектное задание. Вы будете работать с данными об участниках Олимпийских игр за последние 120 лет. Вам предстоит выполнить несколько заданий:

1. Данные разбиты на несколько файлов. Нам надо как-то объединить наблюдения в единую таблицу. Пожалуйста, напишите пользовательскую функцию, благодаря которой мы сможем собрать все наши наблюдения в одну таблицу. Так как олимпиады будут проводиться регулярно, то функция должна объединять все файлы определенного расширения из заданной папки (тип расширения передается как аргумент функции). (5 баллов)

2. Посмотрите, действительно ли все данные корректны? Если найдете, что что-то не так, то исправьте это, пожалуйста. Объясните, почему вы воспользовались именно этим подходом. Может быть у него есть альтернативы? (5 баллов)

3. Выясните возраст самых молодых спортсменов обоих полов на Олимпиаде 1992 года. (2 балла)

4. Рассчитайте среднее значение и стандартное отклонение переменной ***Height*** для спортсменов каждого пола. (2 балла)

5. Рассчитайте среднее значение и стандартное отклонение переменной ***Height*** у теннисисток (sex = F) на Олимпиаде 2000 года. Округлите ответ до первого знака после точки. (2 балла)

6. В каком виде спорта участвовал самый тяжеловесный атлет на Олимпиаде 2006 года? (2 балла)

7. Какое количество золотых медалей было получено женщинами с 1980 по 2010 года? (2 балла)

8. Сколько раз спортсмен **John Aalberg** участвовал в Олимпийских играх в разные годы? (2 балла)

9. Определите наименее и наиболее представленные (по числу участников) возрастные группы спортсменов на Олимпийских играх 2008 года. Возможные возрастные группы: [15-25), [25-35), [35-45), [45-55]. (4 балла)

10. Насколько изменилось число видов спорта на Олимпиаде 2002 года по сравнению с Олимпийскими играми 1994 года? (2 балла)

11. Выведите для зимней и летней олимпиады отдельно топ 3 стран по каждой из типов медалей (2 балла)

12. Создайте новую переменную ***Height_z_scores*** и сохраните в нее значения переменной ***Height*** после ее стандартизации. (2 балла)

13. **Дополнительно**: Создайте новую переменную ***Height_min_max_scaled*** и сохраните в нее значения переменной ***Height*** после применения к ней min-max нормализации (нужно будет разобраться, как она работает). (2 балла)

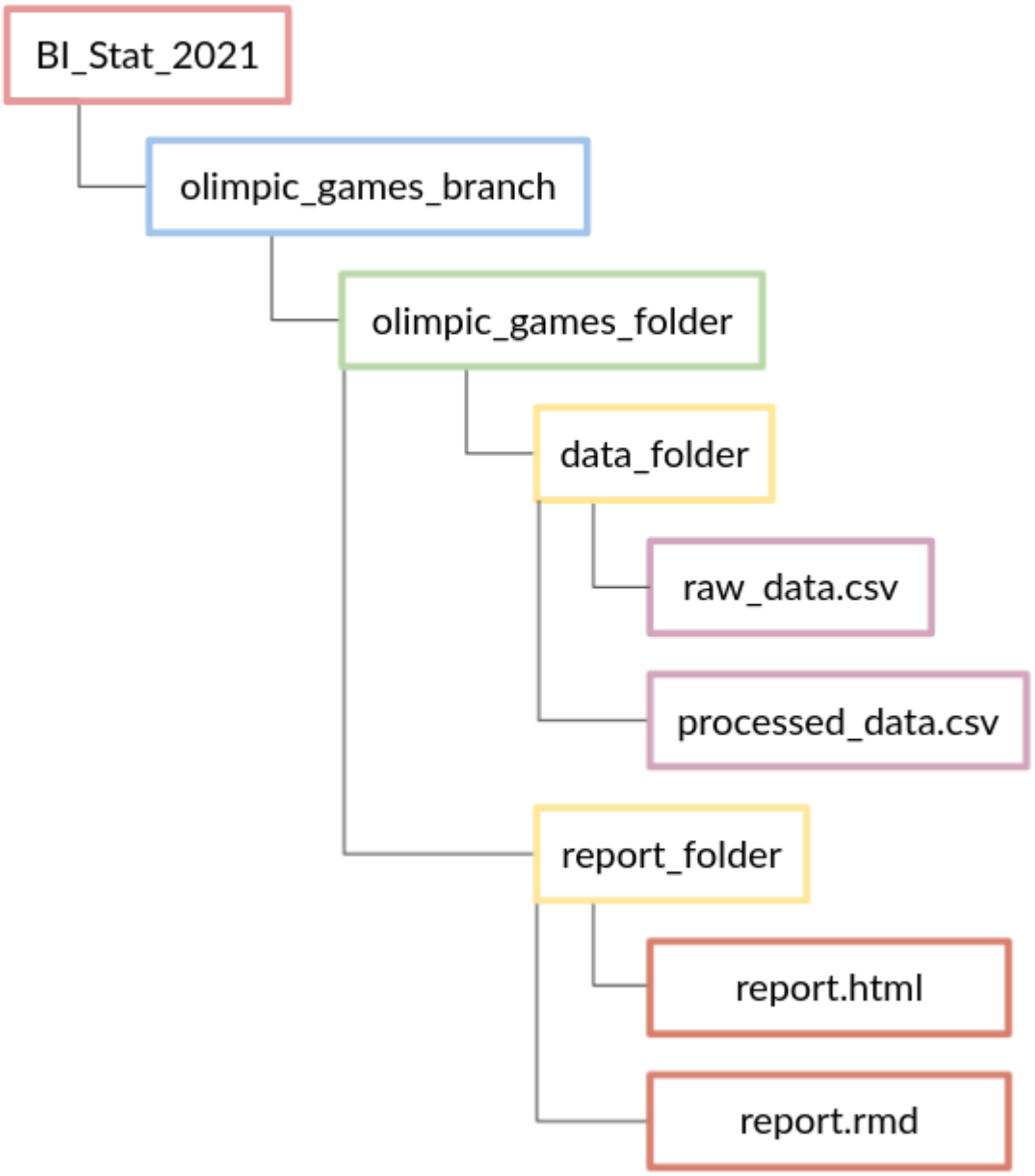
14. Сравните рост, вес и возраст мужчин и женщин, участвовавших в зимних олимпийских играх. Пожалуйста, оформите результаты так, чтобы мы сразу могли использовать их для статьи. (5 баллов)

15. Нас особенно интересуют переменные Team и Medal. Что ты можешь про них сказать? Есть ли у нас основания предполагать, что они могут быть взаимосвязаны? Как ты это определил? (5 баллов)

16. Задание *“со звездочкой”* (**дополнительное**). В нем предполагается выдвинуть и проверить несколько гипотез в рамках тех методов, которые мы освоили (все, что вам покажется интересным). Здесь нет правильного решения, будет оцениваться ваше умение применять изученные методы и интерпретировать полученные результаты. (10 баллов)

Технические требования к отчёту (5 баллов):

1. Отчёт должен быть представлен в формате rmd и скомпилированного html (можно в формате pdf).
2. Вам нужно будет создать репозиторий для нашего курса и назвать его *BI_Stat_2021.* Внутри него создать ветку для текущего проекта. В ветке проекта создайте папку (назовите ее также как и ветку). Внутри этой папки у вас должно быть две директории: data для данных, с которыми вы работаете и report для отчета в формате rmd и html (или pdf). В итоге у вас должно получиться нечто подобное:



Внутри папки olimpic_games_folder вы можете безболезненно создать отдельный файл Readme.md для данного проекта, чтобы коротко описать в нем то, чем вы занимались.

3. Ваш файл rmd должен компилироваться не только у вас на компьютере. За ошибки компиляции баллы будут сильно снижаться.

4. Все разделы отчета, в особенности графики должны быть оформлены в едином стиле. Подписи должны быть полными, логичными и читаемыми.

5. Ваш отчет должен быть универсальным. Так, чтобы при добавлении новых данных он работал корректно. То есть он не должен зависеть от числа файлов в папке и от их названий (гарантируется, что у нужных файлов будет расширение, которое передается в вашу функцию из задания 1)

6. Отчет должен быть структурирован (например, загрузка данных -> краткий EDA -> задачи -> дополнительная часть (по желанию) -> выводы).

7. Ответы на задания должны быть выделены. Когда вы используете тот или иной статистический критерий, то нужно указывать условия его применимости, а в качестве результата приводите значения статистики, p value, число наблюдений и параметры, с которыми вы запускали функцию, если они отличались от параметров по умолчанию.

8. За отчет на английском языке можно получить **дополнительно** 5 баллов.

9. За хорошо оформленный Readme для проекта **дополнительно** 3 балла.

P.S. Итого за отчет можно получить следующие баллы:

Задача 1	5
Задача 2	5
Задача 3	2
Задача 4	2
Задача 5	2
Задача 6	2
Задача 7	2
Задача 8	2
Задача 9	4
Задача 10	2
Задача 11	2
Задача 12	2
Задача 13	2
Задача 14	5
Задача 15	5
Задача 16	10
Структура отчета	5
Отчет на англ.	5
README	3
Сумма	**67**