Задание на обработку данных для Аналитика ЭКОПСИ

Важно: выполняйте задание на языке R. Это основной язык для анализа данных в ЭКОПСИ и мы хотим убедиться, что вы с ним знакомы.

Вводная информация

В папке `data` находятся 43 CSV файла с отзывами на различные вина.
CSV в формате Excel с разделителем полей "`;`", десятичным разделителем "`,`", в кодировке UTF-8-BOM.

Описание данных

В файлах содержатся рейтинги с отзывами (описаниями вкусового профиля) на различные вина. Каждая строка содержит отзыв, оценку, стоимость и информацию о происхождении вина.

переменная	**тип**	**описание**
entry_id	integer	ID записи
country	character	Страна происхождения вина
variety	character	Сорт винограда
winery	character	Винодельня
designation	character	Название виноградника в винодельне
points	integer	Баллы WineEnthusiast на шкале 1-100 (публикуются только отзывы с баллом >=80)
price	integer	Стоимость за бутылку, USD
province	character	Провинция происхождения вина
region	character	Регион в провинции происхождения вина
taster_name	character	Имя автора отзыва
title	character	Заголовок отзыва, как правило содержит винтаж (год урожая)
description	character	Описание вкусового профиля вина от автора отзыва

Образ результата

Вам необходимо провести небольшой анализ этих данных, но для начала очистить их и предобработать.

Задание состоит из двух частей: "техническая" и "творческая".

- У "технической" есть ожидаемый правильный результат.
  - Приложить скрипт, воспроизводящий вашу работу.
  - Вписать ответы в приложенный xlsx файл (`part_1.xlsx`).
  - Вы можете оставлять ответы на вопросы и любою дополнительную информацию в коде вашего скрипта,
    но не забудьте потом **обязательно внести ответы в xlsx файл**.
- Результат "творческой" части полностью зависит от вас.
  - Скрипт, который позволит запустить ваш анализ (**обязательно**)
  - Любые дополнительные материалы, которые нужны вам, чтобы проиллюстрировать анализ (**не обязательно**).

Для обеих частей скрипт может быть одним, нет необходимости его разделять.

Важные требования к скрипту

- Формат - файл `.r` или rmarkdown notebook / quarto / jupyter notebook.
- Кодировка UTF-8.
- Все загруженные библиотеки должны быть перечислены в начале файла.
- Пути до файлов должны быть относительными, напр. `data/reviews_argentina.csv`

Часть 1: Техническая

1.1 Загрузка данных

Загрузите данные из всех CSV в один `dataframe` (или `tibble` / `data.table`).

Вопросы:

1. Какое количество строк в загруженном dataframe?
2. Какое количество столбцов в загруженном dataframe?

1.2 Очистка данных

Данные пришли со следующией дополнительной информацией:

- Строки файлов могли случайным образом дублироваться
- В `taster_name` иногда вместо латинских "a", "o", "e" встречаются их киррилические эквиваленты.

Неизвестно на каких этапах подготовки данных происходили эти ошибки,
сколько раз это происходило и в какой очередности.

Очистите данные от перечисленных ошибок.

Вопросы:

1. Какое количество строк в **очищенном** dataframe?
2. Какое количество уникальных значений содержится в `taster_name` (вкл. `NA`)

1.3 Эксплораторный анализ

Прежде чем переходить к анализу, нужно немного познакомиться с данными.
Ответьте на несколько вопросов о данных, используя очищенный dataframe.

Задания:

1. Для переменных `points` и `price` приведите набор описательных статистик:
    - Среднее
    - Медиана
    - Стандартное отклонение
2. Опишите связь между `points` и `price`
    - Укажите коэффициент корреляции и p value.
    - Укажите, какой коэффициент вы использовали и почему.
    - Если перед расчётом коэффициента вы трансформировали переменные - опишите эти трансформации и их причины.

1.4 Визуализация данных

Для этого раздела необходимо использовать пакет `ggplot2`.
Изображения графиков не нужно копировать в xlsx файл.

Оставьте воспроизводимый код для визуализации и ответьте на некоторые вопросы о вашей визуализации.

Исходите из того, что эти графики будут продемонстрированы аудитории людей, знакомых с предметной областью. Они сами не занимаются анализом данных, но с некоторой периодичностью смотрят подобные отчёты. Графики опрятными и понятными.

Задания:

1. Подготовьте два графика, отражающих распределение переменных `points` и `price`.
    - Какой тип графика вы выбрали для `points` и почему?
    - Какой тип графика вы выбрали для `price` и почему?
2. Подготовьте график, отражающий связь переменных `points` и `price` наиболее наглядным образом.
    - Опишите тип графика (и дополнительные смысловые элементы), который вы выбрали. Почему именно они?

Часть 2: Творческая

В этой части вам не нужно вписывать ответы в excel файл. Но нужно продемонстрировать свой подход к анализу данных.

Предположим, что у этого исследования есть заказчик и он сформулировал свои требования.
Вы не ограничены этими требованиями, но должны их удовлетворить.
Используйте любую удобную для вас форму - текстовый вывод, таблицы, графики и т.д.

Заказчик сформулировал задачи так:

- Как зовут ТОП-5 самых продуктивных авторов отзывов (по числу отзывов).
- Вина из каких стран каждый из ТОП-5 в основном оценивает.
  Не интересны кейсы, если он оценил одно вино откуда-то из необычного места.
- Отличаются ли ценовые диапазон вин, которые оценивает каждый из ТОП-5?
  Хочется увидеть наглядно.

Дополнительно

Выполняйте эту часть, только если у вас есть на это время и ресурсы.
Отсутствие ответа не будет трактоваться как плохой результат.

Заказчику очень интересно узнать - что в вине "работает" на высокий рейтинг (90+).
Интересно, какие есть маркеры высокого рейтинга в описании вкуса и в происхождении вина.