# Тестовое задание Junior Data Analyst

## Часть 1. Работа с данными

Входные данные для тестового задания можно найти [здесь](https://docs.google.com/spreadsheets/d/165sp-lWd1L4qWxggw25DJo_njOCvzdUjAd414NSE8co/edit?usp=sharing) (или [здесь](https://docs.google.com/spreadsheets/d/1SWZCf9MSjP1wNuphLEOH9bA23qbUIyt9/edit?usp=sharing&ouid=100308789753749109392&rtpof=true&sd=true), резервная ссылка).

Ваша задача - подготовить и обработать исходные данных так, чтобы их можно было использовать во второй части задания.

Требования к выходным данным:

1) В выходной таблице должны остаться только следующие колонки:

area, cluster, cluster\_name, keyword, x, y, count, color, где

* area - область,
* cluster - номер кластера,
* cluster\_name - название кластера,
* keyword - словосочетание,
* count - показатель,
* x и y - координаты для диаграммы рассеяния,
* color - цвет точки на карте для данного словосочетания

2) Колонку color нужно добавить самостоятельно - цвета вы можете взять из цветовых палеток [Tableu](https://help.tableau.com/current/pro/desktop/en-us/formatting_create_custom_colors.htm) или по своему усмотрению.

3) Цвет задается каждому словосочетанию согласно следующими правилам:

* внутри одной области цвета словосочетаний в одном кластере должны быть одинаковые, в разных - отличаться (например, у "Кластер 1" все слова будут окрашены в красный, у "Кластер 2" - в зеленый и т.д.)
* цвета кластеров в разных областях могут повторяться
* цвета кластеров в разных областях с разным номером не имеют никакой связи (у одной области [area] слова из "Кластер 1" могут быть красного цвета, в другой области у слов из "Кластер 1" может быть другой цвет)

3) Не должно быть дубликатов слов в одной и той же области (area), но словосочетание может повторяться из area в area

4) Колонки должны называться именно так, как указано в п.1

4) Сортировка должна происходить по колонкам area, cluster, cluster\_name, count (по count значения сортируются в убывающем порядке, в остальных - по возрастающему).

5) Количество переданных в исходных ключевых слов должно совпадать с количество слов в выходных данных (за исключением дублированных строк или строк с пустыми\неформатными значениями по ключевым показателям [перечислены в п. 1], если такие имеются).

6) Никакие другие особенности оформления не должны учитываться при обработке данных (заливка и пр.)

7) Выходные данные должны быть аккуратно оформлены (заголовки закреплены, включен фильтр)

Формат представления выходных данных: google spreadsheet-таблица.

Выполнение данной работы желательно с помощью одной из библиотек:

* data.table ( R )
* pandas (Python)

## Часть 2. Построение графиков

На основании обработанных данных постройте по одной диаграмме рассеяния для каждой области (area) (пример внешнего вида см. в приложенном [svg-файле](https://drive.google.com/file/d/1uxxkyPipGoR4ZLluxV7S1acLZw3Y4VuS/view?usp=sharing)).

Требования к визуализации:

- Наличие Footer-подписи на изображении

- Наличие легенды цветов и кластеров

- Перенос слишком длинных словосочетаний (например, слова длиннее 15 символов, можно разбить на "solar\n cell")

- Минимизация наложения (слепливания) подписей к друг на друга (постарайтесь сделать так, чтобы наложение было минимальным)

- Обводка точек.

Формат представления выходных данных: png-файлы размером не менее 1500х1500 пикселей с визуализациями для каждой области (area).

Выполнение данной работы желательно с помощью одной из библиотек:

* ggplot2 ( R )
* Matploptlib (Python)
* plotly (Python) и т.п.

Требования к результатам:

* Код для первой части задания
* Код для второй части задания
* Таблица с трансформированными данными
* Набор визуализаций

Предпочтительный формат представления кода - GitHub или gist.github, визуализаций и таблиц - Google Docs

# Результат

Ссылку на результаты отправлять по электронной почте [gkuzmin@hse.ru](mailto:gkuzmin@hse.ru) с темой письма “Тестовое задание Junior Data Analyst”