**Проект по дисциплине «Введение в Python для наук о данных»**

1. Разбираться на группы по 2–3 человека (можно индивидуально, но не больше 3-х).
2. Найти и скачать какой-нибудь набор данных (датасет) в виде CSV или Excel-файла. Как вариант, Excel можно сохранить в CSV, а потом этот текстовый файл загрузить на https://pastebin.com. Либо разобраться, как загружать файлы в Google Collaboratory, чтобы загрузить сразу xlsx.
3. В датасете должно быть несколько числовых колонок (это нужно для анализа). Строк должно быть не меньше 100, а в идеале – несколько сотен или тысяч.
4. Сформулировать цель, гипотезы и задачи исследования (применительно к найденным данным). Например, «возраст работника влияет на его зарплату», или «уровень заболеваемости COVID-19 в стране зависит от расходов на здравоохранение в этой стране», или «возраст жителей Пермского края имеет нормальное (или НЕнормальное) распределение», или что-нибудь еще в таком духе. То есть предположить наличие каких-то закономерностей в данных, которые потом будете проверять.
5. Загрузить файл с данными в датафрейм. При необходимости выполнить его первичную трансформацию – удалить нужные колонки/строки, задать индексы строк/имена колонок, преобразовать строковые значения в числовые…
6. Посмотреть на данные, вычислить описательные статистики, визуализировать (построить графики).
7. Выполнить сложный отбор записей в датафрейме (условия для отбора придумать самостоятельно).
8. Проверить данные на наличие пропусков. При необходимости избавиться от них каким-нибудь способом (удалить записи с пропусками; восстановить пропуски).
9. Проверить данные (числовые) на наличие выбросов. При необходимости удалить выбросы.
10. Проверить числовые данные на нормальность распределения (разными способами).
11. При необходимости выполнить нормализацию данных.
12. Построить корреляционную матрицу для числовых столбцов.
13. Для столбцов с предполагаемой (и/или явно обнаруженной с помощью корреляционного анализа) связью построить линейную регрессионную модель.

Все полученные результаты прокомментировать, попытаться дать им интерпретацию, объяснить. Сделать выводы о том, подтвердились ли выдвинутые гипотезы? Предложить направление дальнейших исследований.

Проект оценивается по критериям:

- умение найти и загрузить нужные данные в pandas
- умение выполнять основные операции с датафреймами – добавление строк/столбцов, извлечение данных, отбор, преобразование данных…
- умение выполнять базовые операции анализа данных (выявление и устранение пропусков, выбросов, проверка на нормальность, корреляционный и регрессионный анализ…)
- умение выполнять визуализацию (строить графики).

**Отчет по проекту** представить в виде блокнота в Google Collaboratory (с кодом, графиками, пояснениями).