## Основы работы с pandas

### Цель работы

На практике ознакомиться с основными приемами обработки данных с использованием библиотеки pandas, освоить основные понятия этой библиотеки,

### Методические указания

Для успешного выполнения данной лабораторной работы необходимо заранее ознакомиться с официальной документацией библиотеки pandas.

Работу следует выполнять в интерактивной среде программирования Jupyter notebook. Оценивается полнота выполнения задания, знание и умение применять библиотечные функции

К данной работе прилагается файл с набором данных. Данный набор данных используется для открытого соревнования по машинному обучения Kaggle. Набор описывает основные характеристики пассажиров Титаника с информацией о том, выжил данный пассажир или нет. Часть полей носит категориальный, часть численный характер. Часть данных отсутствует.

### Задания к выполнению

1. Получить файл набора данных, прилагающийся к этой работе [dataset\_1\_Titanic.csv](https://drive.google.com/open?id=16nyTXWRBbiGdI73kS8QIPNbhNICg8inOSeeWsBY4t7o)(60 КБ)
2. Считать файл в pandas DataFrame
3. Ознакомиться с основными параметрами датасета, используя функции библиотеки pandas, включая следующие данные:
   1. Количество точек данных (измерений) в наборе
   2. Количество полей данных (атрибутов)
   3. Тип данных каждого поля, шкала каждого поля
   4. Количество отсутствующих значений для каждого поля
   5. Количество и названия константных и уникальных полей
4. Удалить из набора поля, не несущие информативной нагрузки - содержащие уникальные либо константные значения
5. Заменить возраст пассажиров на год рождения
6. Переименовать все названия колонок на русском языке
7. Заменить все отсутствующие данные средним значением (в случае числового поля) либо самым распространенным в категориальных полях.
8. Сохранить получившийся набор данных в формате CSV
9. Открыть данный файл в программе Microsoft Excel и сохранить в формате .xlsx
10. Считать файл .xlsx в новый DataFrame
11. Повторить п. 2 для этого набора данных

### Контрольные вопросы

1. Какие две главные структуры данных используются в pandas? В чем их отличие?
2. Назовите основные функции чтения данных из файлов
3. Напишите функцию записи набора данных А в файл с разделителями-запятыми
4. Как называется функция, выводящая на экран первые N строчек набора данных?
5. С помощью какой функции можно удалить поле набора данных? А строку?
6. Напишите участок кода, демонстрирующий возможность добавления новой строки к набору данных.
7. Напишите участок кода, демонстрирующий возможность добавления новой колонки (поля) к набору данных.

### Темы для самостоятельного изучения

1. Объединение наборов данных в pandas.
   1. <https://pandas.pydata.org/pandas-docs/stable/comparison_with_sql.html>
   2. <https://pandas.pydata.org/pandas-docs/stable/merging.html>
2. Работа с индексами
3. Группировка данных
   1. <https://pandas.pydata.org/pandas-docs/stable/groupby.html>
4. Работа с датой-временем
   1. <http://strftime.org/>