Семинар 2

Знакомство с Pandas

Аналитика данных

- Суть: анализ таблиц с данными
- Методы анализа на примере Excel: ручной анализ, фильтрация, формулы, макросы, сводные таблицы, диаграммы

В чем проблема при работе с данными в Excel?

- Ваши версии

Пример экселевских данных

- Входящая мобильность персонала и учащихся МИЭМ за 2018 г.
- Какой в этом смысл?
- Сложная структура
- Многогранный анализ
- Неоднозначная визуализация

Анализ табличных данных на Python

Pandas – это быстрый, мощный, гибкий и простой в использовании инструмент для анализа и обработки данных с открытым исходным кодом, созданный на языке программирования Python. На данный момент библиотека Pandas является ключевой в анализе данных (Data Mining).

Основные возможности библиотеки Pandas

- DataFrame - быстрый и эффективный инструмент для манипулирования данными со встроенной индексацией. Методы, требующие высокой производительности, написаны на C или Python.
- Позволяет читать и записывать данные разных форматах: CSV(comma-separeted values), таблицы Excel, базы данных SQL, иерархический формат HDF. Всего насчитывается 19 поддерживаемых форматов.
- Удобный инструмент для работы с отсутствующими данными.
- Простое управление беспорядочными данными в упорядоченной форме.
- Гибкое изменение форм: добавление, удаление, присоединение новых или старых данных.
- Интеллектуальное индексирование, манипулирование и управление столбцами и строками.
- Мощный инструмент для агрегирования и преобразования данных, в том числе и большого размера (BigData).
- Быстрое слияние и объединение наборов данных, например, два и более объектов DataFrame.
- Поддержка иерархического индексирования, то есть возможность объединения столбцов под общей категорией (MultiIndex).
- Поддержка работы с датами и временем.
- Подробнее тут: pandas documentation — pandas 1.4.0 documentation (pydata.org)

Пример работы с Pandas

- Анализ списков групп БИВ185 и БИВ186 в формате csv
- colab.research.google.com/drive/1jgqwp0sadPFXjFIJR4emNumMMNTH9cV0#scrollTo=85RXxEiTIL_e

Задание 1

- Средствами Pandas найти 3 самых часто встречающихся имени в списках групп БИВ 214 и БИВ 215

Задание 2. Анализ Electronic Card Transactions

- Необходимо:

1. Прочитать файл.
2. Определить широту использования типа карт.
3. Перечислить области транзакций.
4. На что тратят больше всего, а на что меньше всего?
5. Посчитать минимальную, максимальную, медианную транзакцию.
6. Визуализировать необходимые (на Ваш взгляд) данные.
7. Сделать краткий вывод по таблице.