**Тестовое задание 1: Анализ финансовых транзакций и клиентов**

**Описание:**

В тестовом задании вам предоставляются два набора данных:

Файл transactions\_data.xlsx (данные о финансовых транзакциях):

* transaction\_id – уникальный идентификатор транзакции
* client\_id – идентификатор клиента, совершившего транзакцию
* transaction\_date – дата проведения транзакции
* service – тип оказанной финансовой услуги
* amount – сумма транзакции (выручка)
* payment\_method – способ оплаты (банковский перевод, кредитная карта и т. д.)
* city – город, в котором была совершена транзакция
* consultant – имя финансового консультанта, оформившего транзакцию

Файл clients\_data.json (данные о клиентах):

* client\_id – уникальный идентификатор клиента
* age – возраст клиента
* gender – пол клиента
* net\_worth – чистая стоимость активов клиента

Данные могут содержать пропуски, ошибки (например, некорректные даты), а также аномальные значения (отрицательные суммы, несуществующие услуги и т. д.), поэтому перед анализом необходимо их очистить и нормализовать.

**Задачи:**

1. Очистка и подготовка данных. Проверить данные на пропущенные и аномальные значения (например, отрицательные суммы транзакций). Удалить или исправить некорректные данные. Привести формат дат к единому стандарту.
2. Анализ данных. Определить топ-5 наиболее популярных услуг по количеству заказов. Рассчитать среднюю сумму транзакций по каждому городу. Определить услугу с наибольшей выручкой (по сумме amount). Вычислить процент транзакций по способам оплаты (наличные, банковский перевод и т. д.). Рассчитать выручку за последний месяц (по сумме amount).
3. Объединение данных. Объединить транзакции с клиентскими данными (по id). Создать новую колонку с уровнями активов. Провести анализ по уровням активов и определить, какие категории клиентов приносят наибольшую выручку. Категории клиентов по уровню активов:
   1. Низкий капитал (<100 000)
   2. Средний капитал (100 000 - 1 000 000)
   3. Высокий капитал (>1 000 000)
4. Визуализация данных (\* дополнительное задание). Построить распределение сумм транзакций. Создать диаграмму выручки по услугам. Построить график зависимости средней суммы транзакции от возраста клиентов.
5. Прогнозирование (\* дополнительное задание). Прогнозирование спроса на следующий месяц (например, с использованием линейной регрессии).

**Формат сдачи:**

* Jupyter Notebook (.ipynb) или Python-скрипт с комментариями,
* Краткое описание шагов и выводов в формате Markdown или текстовом файле.

**Тестовое задание 2 (\* дополнительное задание): Разработка поисковика для данных, представленных в различных документах.**

**Описание:**

Загрузка содержимого офисных документов в базу полнотекстового индекса. Краулинг хранилища файлов doc, docx, xls, xlsx, pdf, а также их вложенностей в архивы (zip, rar, 7z) с последующей загрузкой текстовой информации в FTI-базу.

**Задачи:**

1. Генерация хранилища файлов разных форматов с произвольным содержимым, желательно использовать все указанные в описании форматы, а также архивы.
2. Разработка краулера любым способом (python скрипт, конвейер GNU утилит и так далее) для реализации:

* сканирования хранилища,
* нахождения файлов,
* парсинга типов, названий файлов и главное содержимого,
* сохранения собранных данных в формате csv.

1. Создание базы данных для реализации поиска нужной информации по обработанным краулером файлов. Предполагается использование бд с полнотекстовым поиском: SQLite, MySQL, PostgreSQL. Импорт данных из созданных csv файлов в базу данных

**Формат сдачи:**

* Bash-скрипт, Python-скрипт или другая использованная реализация краулера с комментариями,
* Примеры сгенерированных офисных документов и csv файл итоговой таблицы из базы данных.