**Самостоятельное задание для практики. Занятие 3**

Для выполнения этого домашнего задания вам понадобится **KNIME**. Проверьте, что установили его в предыдущем задании.

**Кейс.** Вы работаете в компании, которая занимается продажами оборудования и прочих товаров в США. Руководитель попросил вас подготовить отчёт по продажам.

Всё, что у вас есть в качестве исходного материала, — хранилище данных, роль которого будет выполнять файл Superstore\_DVA.xls ([ссылка для скачивания](https://drive.google.com/file/d/1ns1UWjjXqrmceCdeQ6Hb-QeOOpKMopBg/view?usp=sharing)).

# Задание

В этом ДЗ нужно подключиться к файлу и посмотреть, насколько в нём качественные данные.

Выполните действия:

1. Откройте KNIME и подключитесь к Excel-файлу. Для этого по аналогии с предыдущим ДЗ добавьте узел Excel Reader и в конфигурации узла выберите скачанный файл Superstore\_DVA.xls. Убедитесь, что в конфигурации в поле Sheet Selection выбрано либо «St first sheet with data», либо «Select sheet with name Заказы». Выполните узел (Execute) и убедитесь, что загорелся зелёный цвет.
2. Для начала исследуйте ваш набор данных на отсутствующие значения. Для этого добавьте узел Extract Missing Value Cause (Manipulation → Column → Transform) и свяжите его с узлом Excel Reader. Выполните и посмотрите на результат (правая кнопка мыши → Extracted Table).

## **Ответьте на вопрос (1)**. Есть ли какие-то проблемы в датасете? Если да, то какие и препятствуют ли они дальнейшему использованию датасета для вашей задачи?

есть несколько строк без индекса - это не критично

есть несколько строк без валовой прибыли, т.к. процент небольшой, то теоретически их можно либо отфильтровать для расчета прибыли, либо уточнить и добавить данные у владельца данных (либо по другим таблицам рассчитать, если они есть в наличии, где есть информация по себестоимости)

Порядка 10 строк не имеют значений в графе индекс. Если принять во внимание, что выручка по этим позициям очень маленькая, считаю что отсутствующие значения не сильно повлияют на результат исследования.

1. Добавьте ещё один узел Excel Reader, из того же файла считайте лист «Возвраты»‎. Выберите его в конфигурации в разделе Sheet selection > Select sheet with name > Возвраты. Выполните узел и убедитесь, что данные были прочитаны успешно и именно с листа «Возвраты».
2. Добавьте узел Joiner (Manipulation > Split & Combine > Joiner). Он позволит объединить таблицу заказов с таблицей возвратов. У него есть два входа для таблиц — два чёрных треугольника слева. Соедините с верхним входом узла Joiner таблицу с заказами (узел Excel Reader Node 1), а с нижним входом — таблицу с возвратами (узел Excel Reader Node 3).
3. Зайдите в конфигурацию узла Joiner. Нажмите на кнопку Add Matching Criterion, чтобы выбрать критерии, по которым осуществите объединение таблиц (джоин). Этим критерием должно быть поле «Код заказа», которое нужно выбрать два раза — и для Top Input, и для Bottom Input.
4. Выполните узел Join. Посмотрите, что получилось после выполнения узла в таблице Join Result, нажав правой кнопкой по узлу Joiner и выбрав соответствующий пункт из выпадающего меню. Снова зайдите в конфигурацию узла и отметьте теперь в разделе Include in Output галочками пункты Matching rows и Right unmatched rows, чтобы получился тип объединения таблиц Right Outer Join. Ещё раз выполните узел Join.
5. После узла Joiner добавьте ещё один узел Extract Missing Value Cause (Manipulation → Column → Transform), выполните его и посмотрите на результат его выполнения.

## **Ответьте на вопрос (2)**. Как бы вы интерпретировали результат? Почему вывелись именно эти значения, в чём там проблемы? Почему не вывелись строки, которые ранее имели отсутствующие значения в столбце «Индекс»‎?

Строки не вывелись т.к. мы использовали Райт Джоин который к правой таблице присоединил все с совпадающим значением Индекс строки из левой.

В два раза увеличилось количество специальных колонок (с 44 до 88).

## **Дополнительный вопрос (3)**. Какие бы проверки вы применили к полям с датами — «Дата заказа», «Дата отгрузки»? Попробуйте каким-либо способом проверить, что в этих полях нет проблем, или же найдите несостыковки.

пропусков по датам в таблице заказов нет, дата отгрузки на 1-7 дней позже даты заказа, пропуски есть в объединенной таблице из-за того, что 2 возврата не имеют аналога в таблице заказов

В этом домашнем задании вы посмотрели на набор данных в KNIME и поработали с проблемой отсутствующих и некорректных значений.

Благодаря возможностям KNIME вы быстро идентифицировали проблему и сделали выводы, что с этим делать дальше.

Проблема некачественных данных на входе очень распространённая, и если бы вы начали работать с этим набором данных не глядя, то дали бы некачественный результат. Поэтому такие базовые проверки проводить на первом этапе нужно всегда.

KNIME позволяет не только провести манипуляции с данными, например, объединить таблицы, отфильтровать их, провести вспомогательные расчёты и т. д., но и экспортировать данные. Например, с помощью узла Excel Writer вы можете сохранить таблицу, которую получили в KNIME, в Excel файл, с которым дальше уже сможете работать в Tableau.

# 