BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

РЕКОМЕНДАЦИИ ПО ОЧИСТКЕ ДАННЫХ С ПОМОЩЬЮ TABLEAU PREP

РЕКОМЕНДАЦИИ ПО ОЧИСТКЕ ДАННЫХ С ПОМОЩЬЮ TABLEAU PREP

- **Подумайте про данные в целом**
- **Узнайте структуру данных**
- **Отслеживайте шаги**
- **Выборочные проверки**

Данные могут генерироваться, записываться и сохраняться в разных структурах, но когда дело доходит до анализа, не все форматы данных созданы одинаковыми.

Подготовка данных – это процесс очистки данных, реструктуризация неверно сформированных данных и объединение нескольких наборов данных для анализа. Она включает в себя преобразование структуры данных, таких как строки и столбцы, и очистка таких вещей, как типы данных и значения. Скорость и эффективность процесса подготовки данных напрямую влияют на время, необходимое для поиска информации. Понимание объема данных, который вы анализируете и его изменения, может ускорить весь процесс.

• Подумайте про данные в целом

Прежде чем вы начнете, важно подумать о том, как пользователи будут использовать данные, которые вы готовите. Понимание этого поможет вам определить, какие данные будут использоваться, сколько данных нужно внести в инструмент подготовки данных и как в конечном итоге структурировать и сформировать данные. Чтобы начать работу, вам нужно ответить на некоторые основные вопросы:

1. Кто будет проводить анализ?
Рассмотрим конечных пользователей данных. Например, на сколько вы единственный пользователь, который будет получать доступ и понимать все данные? Или будет еще кто-нибудь, например, менеджер по маркетингу, который должен понимать, как работает конкретная кампания на основе выбранных, идентифицированных показателей? Если он последний, тогда вам, скорее всего, захочется обрезать набор данных только этими метриками.

Или, может быть, есть код продукта в одной таблице, но менеджер по маркетингу должен знать название продукта. В этом случае вы должны присоединиться к таблицам данных и фактов, чтобы получить информацию. Аудитория имеет решающее значение при подготовке данных, подобно тому, как вы создаете dashboard.

2. Какие вопросы необходимо задать или какие ответы получить?
В процессе подготовки данных важно понять, как люди будут использовать окончательный результат: для комплексного анализа или для быстрой сводки. Эта деталь значительно влияет на процесс подготовки данных, определяя, как объем работы, так и детали.

Обычно вы можете прогнозировать наиболее распространенные вопросы, ответы на которые люди будут искать в данных, исходя из вашего понимания стратегических приоритетов бизнеса, но, скорее всего, появятся так же и непредвиденные. Когда вы готовите набор данных, есть баланс между обеспечением

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

срочных вопросов и возможностью дальнейшего изучения. Например, кто-то может видеть тренд продаж в течение последних шести месяцев, но углубление в суть, на протяжении конкретной недели требует более глубокого анализа и ежедневной детализации данных.

3.Где находятся данные?
Когда дело доходит до этого вопроса, здесь есть некоторые нюансы. Например, есть ли у вас права на доступ к данному источнику данных и есть ли он в правильной форме? Другими словами, когда вы соединяете его с Tableau, можете ли вы провести анализ, который хотите? Вам нужно будет решить оба этих вопроса, прежде чем начать процесс подготовки.

После того как вы сможете получить доступ к необходимым данным, вам нужно определить, где все это происходит. Спросите себя: где находятся данные, в одной таблице или в нескольких таблицах в одной базе? Возможно, вам понадобится объединить несколько баз данных, чтобы добраться до сути ваших вопросов, или, если вам нужно более надежное представление, вам может потребоваться привлечь внешний источник данных. Например, вы можете проанализировать результаты тестов учащихся в вашем районе и хотите посмотреть, как на них влияют социально-экономические статусы, поэтому вы добавляете данные переписи. Часто бывает необходимо вытащить внешние источники данных, чтобы получить полную картину.

- # Узнайте структуру данных

Теперь, когда вы понимаете, как будут использоваться данные, кто будет их использовать и где они находятся, важно понять, как они построены. Вы никогда бы не сделали ремонт у себя дома, не зная, где именно находятся несущие стены. Аналогичным образом, вы не хотите начинать подготовку данных, не зная, какие поля являются зависимыми или связанными друг с другом, как данные были введены (например, вручную или автоматически) или уровень детализации. Знание структуры данных позволяет вам разработать проект перед тем, как двигаться вперед в процессе подготовки данных.

1. Понимание того, что вы ищите.
Прежде чем вводить данные в инструмент их подготовки, важно понять, с чем вы работаете; вам нужно знать, просматриваете ли вы весь набор данных или только их часть. Вам также может понадобиться провести некоторую «разведку», прежде чем вы начнете очистку данных.

2. Определите размер выборки.
Когда вы подключаетесь к большому набору данных, вы, вероятно, захотите ограничить его выборкой, чтобы увеличить скорость процесса подготовки данных и оптимизировать производительность. Могут быть моменты, когда вы хотите увидеть полный набор данных, с Tableau Prep это также возможно. Если выборка не поможет вам решить задачу подготовки данных, вот несколько вещей, которые вы можете попробовать:

- Увеличьте размер выборки данных. Вернитесь на шаг ввода и отрегулируйте количество строк для образца. Вы можете увеличить количество строк или включить все данные, но помните, что это может замедлить производительность. Еще одно предостережение заключается в том, что использование фиксированного количества строк возвращает то, что начальная база данных использует в качестве критериев для быстрого способа возврата запрошенных строк (т. е. это не обязательно означает верхние 1000 строк в базе данных).

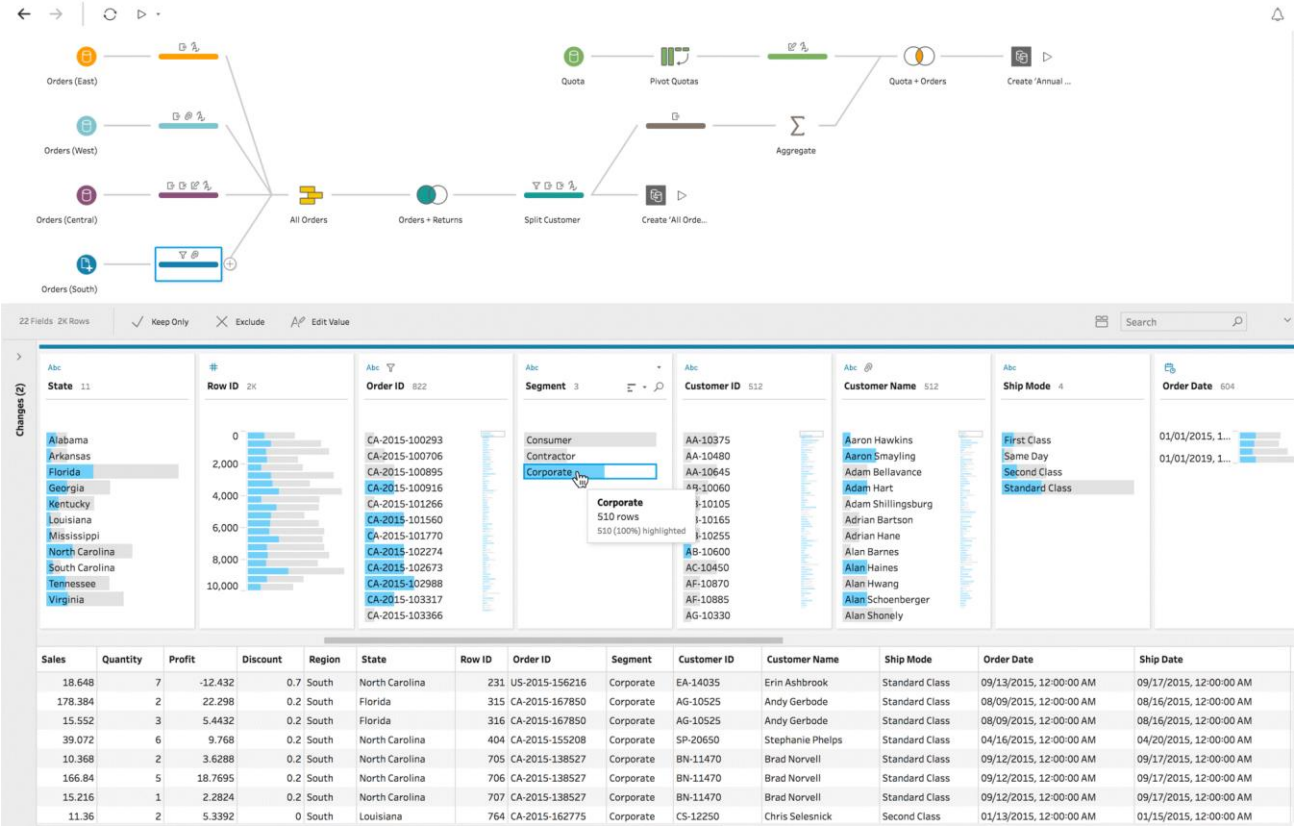
BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

- Используйте случайную выборку. По умолчанию Tableau Prep рассчитает оптимальное количество строк для показа в зависимости от общего количества полей в наборе и типов данных полей. Случайная выборка происходит на уровне базы данных, показывая количество запрошенных строк. База данных просматривает каждую строку и возвращает образец. Эта опция недоступна для всех источников данных и может также влиять на производительность.
- Добавьте фильтр ввода. Добавляя фильтр на этапе ввода, вы гарантируете, что данные, которые вставляются в ваш набор данных, имеют отношение к вашему анализу. Это дает вам более репрезентативную выборку, а также помогает с производительностью.

3. Изучите данные

Во-первых, вы, скорее всего, захотите увидеть уникальное количество значений в заданном поле. В приведенном ниже примере сперва вы увидите в верхней части заголовка столбца количество штатов, представленных в наборе данных. Вы также захотите узнать, как связаны разные значения, чтобы была возможность выявлять отличия или проблемы с данными. В приложении Tableau Prep вы можете использовать подсветку для обнаружения связей между полями. Когда вы нажимаете на значение в панели профиля, оно сужает представление сетки данных, чтобы отображать записи, которые имеют эти значения в указанном поле. Tableau Prep подсвечивает значения полей и окрашивает связанные значения синим цветом.



Sales	Quantity	Profit	Discount	Region	State	Row ID	Order ID	Segment	Customer ID	Customer Name	Ship Mode	Order Date	Ship Date
18.648	7	-12.432	0.7	South	North Carolina	231	US-2015-156216	Corporate	EA-14035	Erin Ashbrook	Standard Class	09/13/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
178.384	2	22.298	0.2	South	Florida	315	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
15.552	3	5.4432	0.2	South	Florida	316	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
39.072	6	9.768	0.2	South	North Carolina	404	CA-2015-155208	Corporate	SP-20650	Stephanie Phelps	Standard Class	04/16/2015, 12:00:00 AM	04/20/2015, 12:00:00 AM
10.368	2	3.6288	0.2	South	North Carolina	705	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
166.84	5	18.7695	0.2	South	North Carolina	706	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
15.216	1	2.2824	0.2	South	North Carolina	707	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
11.36	2	5.3392	0	South	Louisiana	764	CA-2015-162775	Corporate	CS-12250	Chris Selesnick	Second Class	01/13/2015, 12:00:00 AM	01/15/2015, 12:00:00 AM

Синий цвет показывает распределение отношения между выбранным значением и значениями в других полях.

BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

4. Удалите ненужные данные

Чтобы оптимизировать общую производительность процесса подготовки данных, ограничьте поля, которые вставляете в Tableau Prep, теми, которые вам понадобятся для анализа.

Предположим, вы готовите набор данных, который представляет информацию о продажах и продуктах вашей компании. Вы знаете, что позже вы вставите этот набор данных в Tableau для анализа эффективности продаж за год. В этом случае вам может не понадобиться включать подробные сведения о дате отправки для каждого продукта, потому что он не расскажет вам о продаже или почему клиент купил продукт. Это всего лишь показатель того, когда продукт покинул склад, поэтому вы можете удалить его из источника данных. Если в момент подготовки появляется поле, которое вам кажется уже не понадобится, просто удалите его во время процедуры.

Подсказка. В процессе подготовки данных, вы также можете начинать разделять поля, разбивая их на несколько столбцов. После разделения, исходный столбец имеет смысл удалить за ненадобностью.
Фильтрация ваших данных также экономит время в процессе и гарантирует правильность анализа. Например, если вы знаете, что вам нужно всего лишь просмотреть данные о продажах за последние два года, отфильтруйте поле даты на этот временной интервал с фильтром диапазона или относительной даты. Так же могут появится несоответствующие или неверные данные, которые вы захотите удалить. Вы можете просто щелкнуть по значению в области данных и исключить его. Это можно сделать в любой момент подготовки данных.

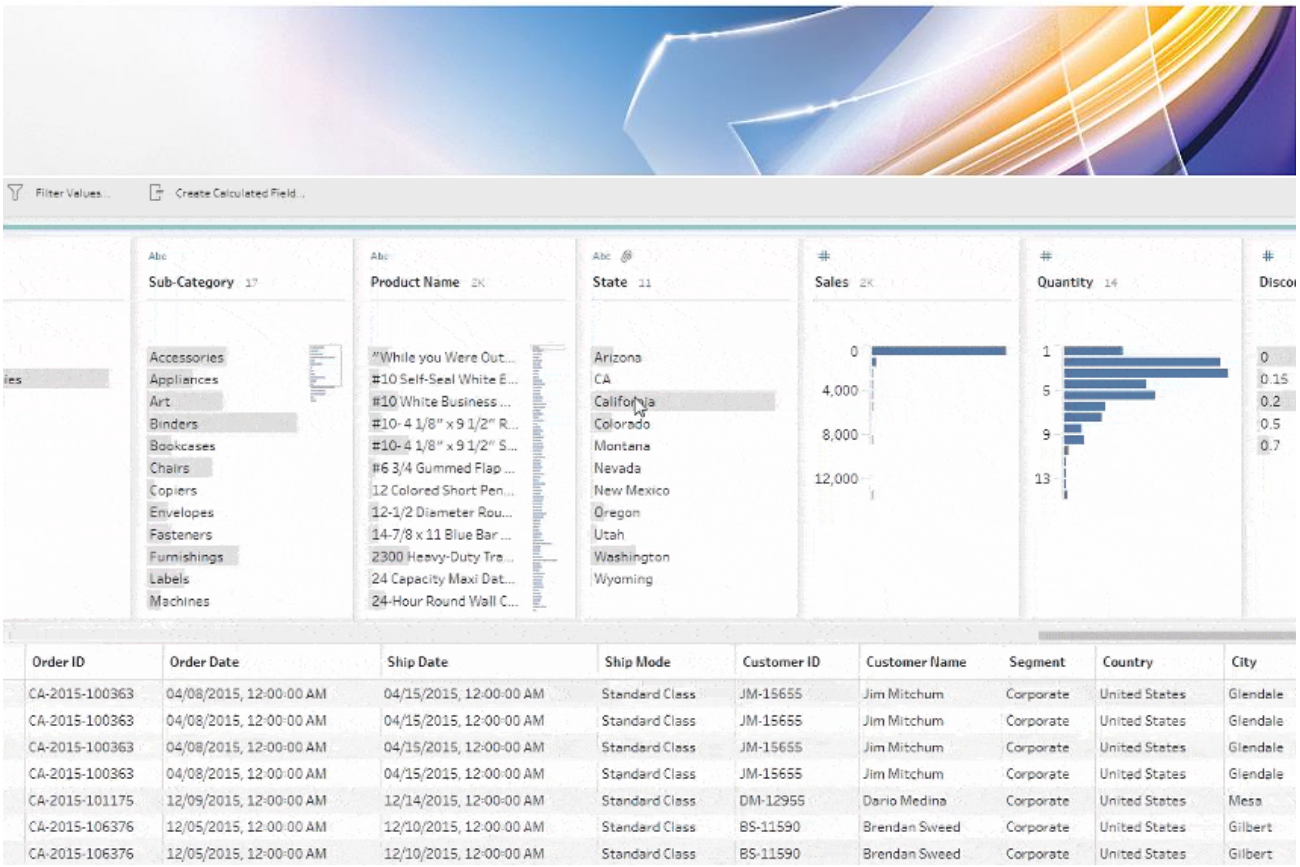
5. Пересмотр и очистка

В Tableau на ваш анализ будут влиять типы данных – и важно правильно определить каждое поле перед началом. Хотя вы и можете редактировать псевдонимы, изменять типы данных, разделять поля и выполнять вычисления в таблице, намного проще выполнить эти действия заранее, особенно при создании набора данных для кого-то другого.

Важно понимать качество данных для каждого поля. Например, номера телефонов, собранные в опросе, могут иметь различные форматы, особенно если они были из глобального пула людей. Вручную проверять тысячи и миллионы уникальных ценностей ради согласованности часто утомительно и подвержено ошибкам. Идентификация шаблонов и обновление данных одним разом имеет преимущество для более чистого набора данных. Так же позволяет сохранить время использование встроенных функций Tableau Prep, таких как «quick clean steps», которые позволяют удалить знаки препинания, цифры, буквы или пробелы.

BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru



Например, если в вашем поле «Штат» есть «California» и «CA», когда остальные значения имеют полные имена штатов, вы можете напрямую изменить значения и «вуаля», теперь записи «California» включают экземпляры «CA».

Вы также можете заметить, что ваши данные в основном верны за некоторыми исключениями. Tableau Prep – умно. Оно помогает стандартизировать значения данных, используя алгоритмы для выполнения тяжелой работы. Возможно, у вас есть столбец данных, где клиенты вводят название своего города. После быстрой прокрутки колонки вы заметите, что в городе Альбукерке есть несколько орфографических ошибок. Вместо того, чтобы обновлять каждый из них вручную, Tableau Prep имеет встроенные функции для группировки и замены обычными символами или произношением. Эти опции используют алгоритмы для упрощения процесса очистки, так что Вам этого делать не придется. Или, если вы упреждаете недостающее значение, вы можете вручную добавить его для включения при запуске всего набора данных через поток. Если вы знаете, что поле должно быть очищено или отфильтровано, но требует чего-то сверх того, что находится в пользовательском интерфейсе, вы можете использовать расчет.

6. Определите конечный результат данных.

Когда вы начнете подготовку данных, возможно будет сложно определить, что должно быть на выходе. Скорее всего, вам понадобится объединить несколько источников данных вместе или развернуть данные из столбцов в строки, чтобы Tableau мог правильно оценить их.

Один из способов решения данной проблемы – представить, как должна выглядеть панель данных в Tableau Desktop.

Если вам нужно объединить две таблицы, у вас есть выбор в использовании т.н. “Join”, или “Union” для данных. “Join” позволяет добавлять в ваш источник данных больше полей, расширяя их количество. Хотя

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

можно добавлять “join” в любое время во время подготовки данных, чем раньше вы его примените, тем быстрее вы увидите общую картину и увидите области, которые требуют немедленного внимания.

Аналогично, “union” позволит вам сложить два набора данных вместе. Например, у вас может быть файл Excel, в котором на каждом листе отображаются транзакции за разные годы. Вместо объединения таблиц “union” позволяет сохранить ту же структуру, но с большим количеством строк.

При ”join” или “union” двух таблиц обратите внимание на уровень детализации. Чтобы правильно объединить их, вам может потребоваться изменить уровень детализации.

- # Отслеживайте шаги

Организация в течение всего процесса подготовки, имеет важное значение, когда вам нужно пересмотреть и внести изменения в какой-то шаг в этом процессе. Хотя вам не нужно следовать определенному набору инструкций по очистке ваших данных (в любом случае, вы должны подготовить данные таким образом, который имеет смысл для вас), процесс подготовки данных будет намного проще редактировать и обновлять, если вы знаете, где вы вносили изменения.

1. Подготовьте данные в ваш способ.
Подготовка данных имеет множество различных компонентов: от перестройки до переформатирования и очистки; вы не должны ограничиваться определенным порядком. Tableau Prep позволяет вам вносить изменения и обновления в данные таким образом, который подходит вам. В то же время некоторые люди могут начать сначала с поворота данных, другие могут начать с очистки орфографических ошибок или отсутствующих данных.

2. Разделение каждого этапа.
Создавая новые шаги для определенного набора действий делает ваш поток приятным и аккуратным. Подумайте о своих шагах как папках в вашем шкафу – вы упорядочиваете файлы по назначению, что облегчает будущий поиск. Аналогичным образом, шаги в потоке должны группировать набор изменений, которые относятся к конкретной задаче. Например, очистка имен клиентов может включать разделение поля, перенаправление связок значений и применение фильтров в других полях для получения правильной сегментации клиента для вывода источника данных. Когда вы выполняете эти действия на одном и том же шаге, вы можете добавить описание, которое поможет вам понять поток позже. Это не только поможет вам, но если вы делитесь потоком с другими аналитиками, это позволяет им находить и ссылаться на одни и те же действия, предоставляя им способ легко сделать любые изменения.

BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

Следить за тем, что произошло на каждом шаге, легко. Если вы совершили действие, которое хотите изменить, вы можете это сделать в области изменений.

- # Выборочные проверки

Важно, чтобы вы понимали, что происходит с данными, когда вы чистите и вносите в них изменения. Вы не хотите слишком углубляться, чтобы понять, что вы присоединились к неправильным двум полям. Это возвращает нас к вопросу о понимании данных. Если у вас есть хорошее представление о том, как должны выглядеть данные, эти выборочные проверки будут легче распознать, когда что-то пошло не так.

1. Использование визуальной обратной связи

Гораздо легче подготовить данные, если вы можете увидеть, как они связаны, прежде чем начинать анализ – например, количество строк в наборе после присоединения таблиц или орфографические ошибки. Как и Tableau Desktop, Tableau Prep был построен с учетом нашей миссии: помочь людям увидеть и понять их данные.

«Сетка»

Использование сетки данных в Tableau Prep идеально подходит для поиска. Вы можете увидеть, как выглядят данные после внесения изменений и получить представление о существующих аномалиях.

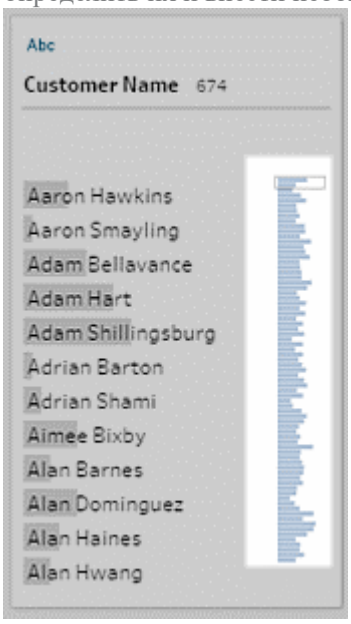
BI CONSULT

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru

Order Date	Region	Ship Date	Row ID	Order ID	Ship Mode	Customer ID
11/22/2015	Central	11/26/2015	15	US-2015-118983	Standard Class	HP-14815
11/22/2015	Central	11/26/2015	16	US-2015-118983	Standard Class	HP-14815
11/11/2014	Central	11/18/2014	17	CA-2014-105893	Standard Class	PK-19075
12/09/2016	Central	12/13/2016	22	CA-2016-137330	Standard Class	KB-16585
12/09/2016	Central	12/13/2016	23	CA-2016-137330	Standard Class	KB-16585
10/19/2017	Central	10/23/2017	35	CA-2017-107727	Second Class	MA-17560

«Мини-карты»

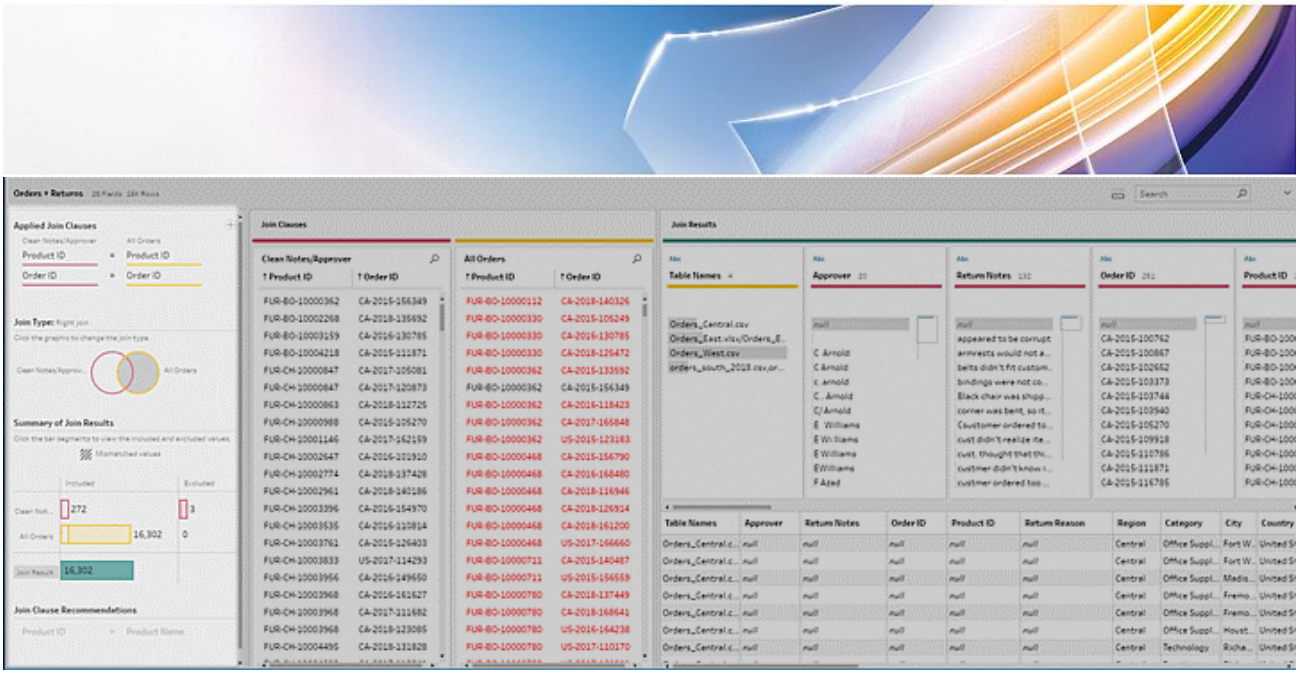
Бывают моменты, когда вы считаете, что данные вроде как уже готовы к анализу, но используя мини-карту вы замечаете аномалию или несколько пропущенных записей. Используйте мини-карту, чтобы определить их и внести необходимые изменения.



«Понять соединение»

Существует вероятность случайного соединения не тех полей, особенно когда вы присоединяетесь к нескольким полям. Tableau Prep использует визуальную обратную связь, чтобы показать вам результаты соединения, помогая вам узнать, есть ли какие-либо отклонения, количество показанных данных, а также их корректность.

BI Consult
199048, Санкт-Петербург,
6 линия В.О., д.63/1, 4 этаж
Тел./факс +7(812) 334-08-01
E-mail: info@biconsult.ru



2. Продолжение обновления

Подготовка данных – это непрерывный процесс. Он не заканчивается, как только вы исправили все орфографические ошибки или объединения. Когда набор данных обновится, ваши вопросы могут измениться или вы обнаружите, что вам нужно добавить другое поле. С помощью функции «Open sample in Tableau Desktop» легко и без проблем можно проверить, как данные отображаются в строке анализа.