Tableau Prep Guide

Tableau Prep is a tool used to build relational databases, manipulate and clean data, and ultimately organize and modify a collection of disparate datasets into a usable format. As you use Prep you will be building a “data flow” that will detail at each step what changes are made. In this guide we will be showing an overview of a number of features in Prep, but by no means are exhausting their potential. When you are manipulating the client’s data, make sure you have a clear understanding of what your end goal is, and what steps you are using to get there. Joining the data at incorrect fields, or mistyping a calculation can misrepresent the data, which is a serious mistake. When in doubt, ask or write down your logic.

This guide will use a collection of lemur data, which we have kindly been given permission to use by Meredith Semel and the Global Change Center at VT. The two tables can be found here:

- Feeding Data
- Behavioral Data
- Ethogram and Key (has the explanation for most abbreviations used within the data)
- This data is broken into 11 sites, each a different location in Madagascar where the data was collected. The feeding data focuses on lemur feeding preferences, frequency, and amount, along with data about the trees feed on. The behavioral data includes observations about what the lemurs were seen doing throughout the day.

In the guide below each step will begin with an explanation. This will include the goal of the step, along with the logic behind why the steps are carried out. If you are confused as to why you are carrying out a certain step [Google or ask for an explanation].

Table of Contents:

- Connect to Datasets
- Cleaning the Datasets
- Joining the Datasets
- Output Data
- Wrap Up
  - Related articles

Connect to Datasets

**Explanation**

To begin, our first goal is connecting our data to Tableau Prep to clean and manipulate it. Prep can take a wide number of different data types, along with the ability to connect to servers directly for live data. We will be focusing on using Excel sheets. These, along with CSV (comma separated values) are the most common types of files we will be interacting with.

Once you have connected the data you will see that each Excel file has multiple sheets within. Each sheet will be treated as a unique dataset, which is why we union these sheets in the beginning.

A WildCard Union will automatically union the datasets contained within the Excel file. This is not something that will be applicable to every file. It is done in this case because each sheet is a different site, all which have the same column headers and can be easily joined.

We will be using Unions. Follow and read the link if you are unfamiliar with them.

1. Using the blue bar on the left, click Connections then Microsoft Excel
   a. Connections, Microsoft Excel
2. Navigate to, and open, the files "behavioral_data.xlsx" and "feeding_data.xlsx"
   a. If files are in the same folder you can select and open multiple files at once
3. Beginning with Feeding_Data, drag the first dataset onto the Prep worksheet.
   a. In the Input window that appears click "Wildcard Union" and then "Apply"
   i. **Wildcard Union, Apply**

4. Repeat Step C with Behavioral_Data
Cleaning the Datasets

**Explanation**

Our first goal is to convert Latitude and Longitude from Degrees Minutes (S13 15.583') to Degrees (-13.25), which is the format Tableau can read coordinates in. In Degrees Minutes the first number is the degree (13) and the second is the Minutes (15.583).

In the calculated field for Latitude we multiply the results by -1 because these are all in the Southern hemisphere, which is displayed as a negative degree for latitude. As longitude is East we keep those values positive.

Our second goal is to create a field that takes information from both "DATE" and "TIME" in the Behavioral data. Tableau expects data and time to be reported in one field, and when that is not present it will default to the first available time (12:00AM) and date (12/30/1899).

We will be using Calculated Fields. Follow and read the link if you are unfamiliar with them. They allow for quick calculations ([field 1] + [field]) or manipulations. Calculated Fields require the most basic knowledge of programing syntax, but should be approachable by all.

1. The following Steps are done with the Feeding Data.
   a. Add a Step to the dataset by clicking the plus sign to the right of the dataset, then Add Step...
b. Click on the step you just created

c. Rename "LAT" to "Latitude Original" and "LONG" to "Longitude Original"
   i. This is done so another user can quickly tell what the fields are without having to spend time reading the data
   ii. Right click on the ID field and click "Rename Field"

   1. Right Click field, Rename Field

   ![Image of data field renaming]

   d. Split Latitude Original and Longitude Original
      i. This is done so we can do the calculations necessary on the numbers within the field, and will strip out the other characters.
      ii. Right click on the Latitude Original and Longitude Original Field, hover over "Split Values," and click "Automatic Split."

      1. Right click field, Split Values, Automatic Split

   ![Image of data field splitting]

      2. NOTE 1: For Longitude Original there was only one degree (49) in all data, so it was discarded in the split. This is the reason our calculations below have minor variations.

      3. NOTE 2: The split fields were renamed the following to be more descriptive:
         a. "Latitude Original - Split 1" to "Latitude Degrees"
         b. "Latitude Original - Split 2" to "Latitude Minutes"
         c. Longitude Original - Split 1" to "Longitude Minutes"

   e. Convert Latitude and Longitude split fields to "Number (decimal)"
      i. This is done so we can carry out calculations on the numbers
      ii. Click on the "Abc" icon to the top left of the field and click "Number (decimal)."
f. Create a Calculated field

i. This is the actual calculations done on the data to determine the Degrees, this will be done once for Latitude and once for Longitude

ii. Click on “Create Calculated Field” at the top of your data panel

iii. Paste the syntax below into the window that appears and name the calculation either “Latitude” or “Longitude”.

1. For Latitude:

   Latitude

   

   $\left(\left[\text{Latitude Degrees}\right] + \left(\left[\text{Latitude Minutes}\right]/60\right)\right) \times (-1)$

2. For Longitude:

   Longitude

   

   $\left(49 + \left(\left[\text{Longitude Minutes}\right]/60\right)\right)$

   a. NOTE: 49 was the constant latitude degree discarded during the automatic split, so it is re-added here as a constant in the calculation. If there was not a constant degree value mimic the formula for Latitude.
2. The following steps are done with the behavioral data (and will be less detailed than the steps above for repeated steps):
   a. Add a step from Behavioral Data
   b. Create a Calculated Field and paste the following syntax, along with naming the field "Behavioral Date Time":

   ```
   Behavioral Date Time
   DATE TIME(STR([DATE]) + " " + STR(DATEPART('hour',[TIME])) + ":" + STR(DATEPART('minute',[TIME])) + ":" + STR(DATEPART('second',[TIME])))
   ```

c. Remove the DATE and Time fields
   
   i. This is done because we have used those fields to create a cleaner field above. This data is now redundant, and unnecessary.
   ii. Right click on the DATE and Time fields and click "Remove Field"

   1. Right click field, Remove Field

   iii. NOTE: removing a field can be easily undone by deleting that step from the changes panel. This panel is on the left side of the grey data window. This is only if you realize you would like to keep the data deleted, and will not be needed for this tutorial. You can also use this panel to rearrange the chronological order the changes occur. This is useful if you forget to change the field type to a Number before creating a calculated field. You can change the field type, then drag the calculated field step below the change, which will then solve the error.
   1. Click on "Changes" to expand the panel
Joining the Datasets

Explanation

Now that we have cleaned the Feeding Data, we want to join it with the Behavioral data to produce one dataset that has both components. We will be joining per site, meaning at the end we'll be able to see for each site how the lemurs behaved and what they ate.

In joins, mismatched values are those that do not have a matching pair in the second dataset. Being a mismatched value does not immediately exclude that portion of data’s usefulness; it simply means it is not further explained in the additional dataset. For example, our feeding dataset includes information about the group “bek_av,” while the behavioral dataset does not. We do not want to exclude this feeding data solely on the basis that it doesn't have complimentary data, so we will include it in the final dataset. When using joins, think about what you are trying to accomplish, look at what may be excluded, and evaluate how the resulting dataset will change your data.

We will be using Joins. Follow and read the link if you are unfamiliar with them.

1. Drag and hover the Behavioral Data step over the step from Feeding Data. On the right side of the Feeding Data step an option will appear titled "New Join." Hover over that option with the Behavioral Data step and let go of the mouse. You should now have a join.
   a. Tableau should have automatically joined the sets using "Table Names" as the Join Clause.
1. a. We want to join these datasets by group, to ensure the data for each lemur group is properly combined. To do this we need to change the Join Clauses to "Group." Click each join clauses and select "Group" on both.

b. We want to join these datasets by group, to ensure the data for each lemur group is properly combined. To do this we need to change the Join Clauses to "Group." Click each join clauses and select "Group" on both.

i. Change join clauses to “Group”
1. We now need to review our join and correct any errors that arise. To the right of the join summary you should see the values within each dataset that were used for the Join Clause (in this case, Group). The mismatched values, in red, are "bek_av," "null," and "6." In the steps below we will modify our data somewhat to result in the best resulting dataset.

ii. Including "bek_av":
   1. As discussed earlier, the "bek_av" group has feeding data, but no behavioral. The feeding data is still useful to us, and we should include it. To do so, change the Join Type to a **Left Join** by clicking inside the blue circle in the Join window.
1. **Click inside blue circle**

![Diagram showing feeding data and behavioral data](image)

- **Applied Join Clauses**
  - `GROUP` = `GROUP`  
- **Join Type**: Left join
- **Summary of Join Results**
  - Included: 567,678
  - Excluded: 0

2. **You should now see that while “bek_av” is still mismatched, it is now included in the join, resulting in 0 values excluded from the feeding data.**

   - **Renaming “6”:**
     1. One of the groups in the behavioral data is named “6,” which does not have a counterpart in the feeding data. Comparing the two sets, it logically seems like “6” should join with “antsaharaingy_6.” We do see that “antsaharaingy_6” already exists in the behavioral data, so we will need to check if our logic holds.
     2. Going back to the behavioral step (click on the step) we can click on “6” within the group field. This will show highlight in blue on all other fields where these values are connected. Doing this, we see that all “6” values are found within the site “antsaharaingy,” which further backs our thought process.

   ![Excel sheet showing data](image)

   - **Optional Step**
     I also recommend when doing exploratory analysis of yours, or a client’s, data as we are that you review the original datasets as well. I did this by opening the “antsaharaingy_6” sheet in Excel, and saw that half way through the data the entry was changed from “6” to “antsaharaingy_6.” This concretely proves that those group values are from the same dataset. Whenever you are working with someone else’s data it is always paramount that you are certain of your hypotheses before
3. Now that we have proven to ourselves "6" is the same dataset as "antsaharainy_6" we can now edit the value. Within the Group field, right click "6" and click "Edit Value." Edit the value to be "antsaharainy_6" and hit enter. You should see the two values merge and only be left with "antsaharainy_6". If not, check if a spelling error occurred and re-edit the value.
   a. Right click "6", Edit value, "antsaharainy_6"

   ![Image of Group field with options]

   b.  

4. Going back to the Join, we should now see that "6" is gone, having been edited to match "antsaharainy_6", and is no longer a mismatched value.

   ![Image of Join window with data]

   a. Checking "null":

   ![Image of data flow diagram]
1. Looking at our Join Results we see that there are 15 excluded values in the behavioral data, and only one mismatched value remaining: "null." By going back to the behavioral step and clicking "null" in the Group field (like we did with "6"), we can see in the data window that these null values are null for every field, and thus are fine to exclude.

2. We should now feel comfortable that we have properly joined our data and are ready to output the flow as one dataset.

**Output Data**

---

**Explanation**

We are now nearly done with Prep, all that is left is to export the data to use in Tableau Desktop. This step can also be used to export the data as a CSV to use in other programs.

Exporting the data runs all the steps in chronological order in the data flow. Depending on the size of the data set, the steps ran, and the power of your computer this may take a while. If you are see a significant time being spent to run an output I recommend breaking the steps up into individual Tableau Prep files, outputting the results, and joining at the end (for example, making the feeding and behavioral changes in different prep windows, outputting each individually, and joining those outputs together). This will cut down on the level of processing needed at once.

1. Click the + icon to the right of the join and select "Add Output"
   a. **Add Output**

![Diagram of data flow with steps and icons]

b.]

2. Select where you want the file to save, and file type (Tableau Hyper is recommended if you are moving to Tableau Desktop next), and click "Run Flow"
   a. **"Run Flow"**

![Screenshot of output settings]

b. [Add output button]
Wrap Up

You now should have successfully unioned, cleaned, manipulated, and joined the datasets together to create one dataset! Through the steps above you connected 21 datasets together quickly, and are now ready for analysis. On your own data, you may find that a step was done incorrectly or additional manipulation is needed. If so, continue to iterate between Tableau Prep and Desktop until the output is properly cleaned and joined. The Tableau community is full of very committed and talent users, and most questions have been answered on their forums. This is just the start of what Prep can do, and with advanced calculations it can be pushed much farther. If you see any errors or confusing portions in this guide please contact Jonathan Briganti at jonbrig@vt.edu. My final Tableau prep flow and output are included below:

- Tableau Prep Completed Flow
- Lemur Completed Output

Last updated 8/17/2018


Related articles

- Tableau Prep Guide
- Tableau Desktop Guide
- Software Download Locations
- ADS Procedure