**In-Class Exercise #9: ETL using Tableau Prep**

|  |
| --- |
| **Submission Instructions**Submit 1) the final version of your combined data file (“EditedNYCParkingData.csv”) and 2) the answers for questions on page 15 (word document) through the form.  |

In this exercise you will use Tableau Prep to perform extract, transform, load (ETL) tasks to combine data across several datasets. You can do many transformation operations on your data using Tableau Prep, including reconciling data in different formats, filtering data, adding new columns to a data set, splitting a single column into multiple columns, and appending rows to the end of a dataset. The purpose of these actions is to create an analysis-ready data set.

Tableau Prep can use data from a variety of sources, including comma-separated value (CSV) files and MySQL databases.

**Part 1: Connect Tableau Prep to Data for Extraction**

1. Start Tableau Prep on your computer.
2. Click the “Connect to Data” button.

3. Select “MySQL” under the list of data sources (in the “To a Server” section).
4. Fill in the dialog to connect to the class MySQL server. Remember the server name is “dataanalytics.temple.edu”, the port is “3306”. The username and password are the same as what you used for MySQL Workbench. You do not need to check “Require SSL.”



NOTE: If you see the following message:


Click on the ‘download and install the drivers’ link, and follow the instructions to download the drivers to your computer.

1. Click Sign In. You will now see a drop down menu of all the databases that are accessible from your MySQL account. Select the parking database (yours will look a little different – look for the word “parking” in the name).


2. You’ll now see the list of tables contained within the database



**Part 2: Transform Data by Reconciling Different Values**

Tableau Prep uses “flows” as its metaphor for working with data. A flow is a series of steps; each step performs one or more ETL operations on your data. A flow looks a little like a flowchart. This is the flow that you will be building as part of this exercise:



There are two other useful things to keep in mind when you’re working with Tableau Prep:

* Tableau Prep flow files do not contain your data. The flow file just contains the instructions for each step you perform on the data to transform it.
* Tableau Prep does not change your original data sources – it creates a new data set by making changes to a copy of that data and saving it as a new file.

Now get started:

1. Drag the Vehicle Information Table onto the workspace. It should look like this:



1. In the Input pane, click on “Data Sample” and select “Use all Data” to ensure that Prep does not automatically choose a sample of the data.
2. Click on the plus sign next to vehicle\_information on the workspace and select “Add Step.” Add Step gives you a general data cleaning step to allow you to view and edit the data before any other steps are done. You should now see a screen like this:


In this section, you can see the exact number of rows and columns within the Vehicle Information Table, by looking at the **Profile Pane**. For this data set, there are 21,000 rows and 9 fields. The **Profile Card** shows the data type of each column and the options available to change or edit the data in those columns.
3. Look at the second column, called “Registration State.” Currently, the data format is a string (denoted by the “Abc”), and the data role is not specified. In Prep, you can specify data roles to ensure that data is correct. Click on “Abc” in the Registration State column, then select Data Role/Geographic/State/Province:
4. You’ve now told Tableau Prep that the data specifically represents States and Provinces. You can see this change by expanding the Changes pane by clicking on “Changes (1)”:



If you delete the “Change Role” entry the data will revert back. Don’t do this, though! We want Tableau to recognize “Registration\_State” as geographic data.
5. Scroll through the data and observe that Registration State is sometimes represented by the two character abbreviation (PA), and other times by the full name (Pennsylvania). We want all state data to be represented the same way so data from PA and Pennsylvania aren’t considered separate locations.

Hover over the Registration\_State column and you’ll see three dots:



Click on the three dots and you’ll see more options. Select Group and Replace/Manual Selection.

1. Select “NY” in the left side of the “Group and Replace” dialog and “New York” in the right side of the dialog in order to replace all instances of “New York” with “NY.”



When successful you’ll see New York disappear from the right side and a paper clip will appear next to NY on the left side:


2. Repeat these steps for all remaining states that are represented this way (i.e., NJ/New Jersey, PA/Pennsylvania, TN/Tennessee, and VA/Virginia. Then click Done to close the dialog.
3. Scroll to the top and remove “99” from Registration State by right clicking on that entry and selecting Exclude. That will remove all entries with a Registration\_State of 99 from the data set.
4. At this point it would be a good idea to save your work. Choose Save from the File menu and save your Tableau Flow file to your computer.

**Part 3: Transform Data by Splitting Data in Combined Fields**

Another important part of data transformation is making sure a column of data is only about a single thing. In the world of relational databases, this is called “atomicity.” For example, a data column like this can be a problem because first name and last name are combined into a single data element. It is difficult to sort and filter by only first name or last name.

|  |
| --- |
| Name |
| Keaton, Alex |
| Moore, Nick |
| Reed, Ellen |

Therefore, we’re better off splitting this into two columns, using the comma as a separator:

|  |  |
| --- | --- |
| Last Name | First Name |
| Keaton | Alex |
| Moore | Nick |
| Reed | Ellen |

In this section, you’ll be splitting columns to make sure each data element is in its own column.

1. Drag the Violation\_Information table onto the workspace.
2. In the Input pane, click on “Data Sample” and select “Use all Data” to ensure that Prep does not automatically choose a sample of the data.
3. Click on the plus sign next to violation\_information on the workspace and select “Add Step.”
4. Look through the data set in the lower section of the window. Find the Law\_SectionSub\_Divsion column. This column has two values separated by a comma. Each data element should be separated into its own column.
5. Hover over the Law\_SectionSub\_Division column in the middle section of the window. Click on More Options (three dots) and then select Split Values/Custom Split.


6. In the Custom Split dialog, enter a comma as the separator and ‘All’ for “Split off.” Then click the Split button.



Scroll all the way back to the left of the data to see the new columns – they will be the first and second columns and will be called “Law\_SectionSub\_Division - Split 1” and “Law\_SectionSub\_Division – Split 2.”

1. Rename the first column (“Law\_SectionSub\_Division - Split 1”) to Law\_Section by right clicking on the column and selecting Rename Field. The rename the second column (“Law\_SectionSub\_Division - Split 2”) to Sub\_Division.

**Part 4: Transform Data Using a Calculated Field**

Look at the Issue Date column within Clean 2 of Violation Information. The date is written as month/day/year. We want to know what day of the week the violation occurred. This requires converting Issue\_Date to a day of the week, such as Friday. This will enable us to analyze whether more violations of a certain type occur on certain days of the week. To do this, we can use Tableau Prep’s calculated field feature. This allows us to create a new data column based on the data in other columns.

1. Click on the “Abc” at the top of the Issue\_Date column in the middle of the window, and make sure that the Data Type is set to Date & Time.


2. Hover over the Issue\_Date column and click on More Options (three dots) and then selection “Create Calculated Field…”
3. This displays the Add Field dialog. You can choose from many different functions. You will be using the DATENAME() function, which returns part of a date as a string value, such as the day of the week.
4. Replace [Issue\_Date] in the equation area with DATENAME(‘weekday’,[Issue Date]) and enter “Day of Week” for the field name. You should see this:
 Don’t click ‘apply’

If you did it correctly you’ll see “Calculation is valid” in the bottom right of the window.
5. Click ‘Save’ and Tableau Prep will create a new column called Day of Week as the first column in your data set. Scroll back to the first column in your data set at the bottom of the window to see this.

**Part 4: Transform Data by Combining Several Sources**

Now that our data is clean, we want to combine several sources into a single data set. In order to do this we will use Tableau Prep’s join feature. As with SQL, a join is a method of combining data by matching on a common field.

1. Select the “Clean 2” operation in the Tableau workspace to see the contents of the violation\_information table. Look at the Description column – this is supposed to contain the descriptions of the violations in the Violation\_code column. All of the values are missing, which makes it hard to easily identify the violations that people commit.
2. Drag violation\_description table to the workspace.
3. Click the plus sign next to the violation\_scription table and select “Add Join.”
4. Drag the cleaned Violation Information (“Clean 2”) to “Add” section of the newly created join.



You should now see this (colors may be different; they are randomly assigned):

5. Click the “Add” button under “Applied Join Clauses.” Then select the Violation\_Code fields from the violation\_description and Clean 2 data sources.


6. The data from Violation Description has now been combined with Violation Information.
7. Add the location\_information and issuer tables to the workspace.
8. Next to Join 1 click the plus sign and select “Add Join.” Drag the Clean 1 data source from vehicle\_information over Join 2 that you just created, making sure “Add” is highlighted as before. You will now see this:



1. Select Join 2 and click the Add button under Applied Join Clauses. Use the “Summons Number” field Apply a Join Step by creating the join clause using the Summons number.



1. Repeat this for the location\_information and issuer tables. You’ll need to create two more joins. Continue to use “Summons Number” as the join field (make sure “Add” is highlighted as before). Once all the joins are completed, you will have this flow:


2. Click on the plus sign next to Join 4 and choose “Add Step.” This will let you browse through all of your data. You can see that violation data, vehicle data, location data, and issuer data are now all part of one table, and the data has been reconciled and transformed.
3. If you haven’t already, now would be a good time to save your work again.

**Part 6: Export Updated Data**

However, keep in mind that Tableau Prep does not alter the original data – it just applies the transformations you’ve specified in your flow diagram to the data. Tableau Prep will generate a new data file based on the actions you’ve specified in your flow.

The data is now combined and transformed. In order to use other software to analyze the data, you can export the transformed data as a Comma Separated Value (CSV) file.

1. Delete the “Clean 3” step you just inserted. You won’t need that. Join 4 will again be the last step in the flow.
2. Click the plus sign after Join 4 and select “Add Output.”
3. In the Output pane, click the Browse button and select the location where you want to save your output file. Name your file NYCParkingData.
4. Choose “Comma Separated Values (.csv)” as the Output type.
5. Click the “Run Flow” button. Once finished you will see this popup.



Click “Done.”

1. Save your Tableau Prep file one more time and exit Tableau Prep.
2. Find your NYCParkingData.csv file and double-click on it to open the file in Excel.
3. Browse the file. Notice that all the columns across the five tables you’ve joined are in the file. Also notice that the transformations are in the file – i.e., the Day of Week column, only two character abbreviations are in the Registraton\_State column, etc.

**Try It on your Own**

1. Open a new MySQL connection and select the parking database.
2. Drag the ticket\_information table onto the canvas and explore the data. This is the second part of the NYC Parking Ticket Data. The data is similar to the data you have just worked on in section one of the exercise, but the format of the data needs to be modified before its records can be appended to the data from NYCParkingData.csv.
3. Open the NYCParkingData.csv file in Excel, and compare the columns to see how the formatting differs between the two data sets. The data needs to be in the same format, including the names of column headings, before it can be combined.
4. Close Excel and return to Tableau Prep. You’ll need to make the following transformations to the data in the ticket\_information table:
* Split the Law\_Section\_Sub\_Division column into two columns and change the new column names to Law\_Section and Sub\_Division
* Change the data type in the Issue\_Date column to “Date & Time”
* Created a calculated field to create a “Day of Week” column
* Separate the Street\_Code1-Street\_Code2-Street\_Code3 column into three separate columns. Change the column names to “Street\_Code1”, “Street\_Code2”, and “Street\_Code3”.
* In the Registration\_State column, change “New York” to “NY” and “Georgia” to “GA”.
* In the Registration\_State column, exclude value “99”.
* Join the violation\_description table with the ticket\_information table to integrate the violation description data into the data set. Use Violation\_Code as the common field in the join clause.
1. Save your work! Save the Tableau Prep flow file to your computer.
2. Connect to the NYCParkingData.csv file by clicking on the plus sign next to Connections and then selecting Text file under “To a File”



Browse for the NYCParkingData.csv file and open it.
3. You’re now ready to integrate the data from the NYCParkingData.csv file with the data you’ve cleaned and joined from the ticket\_information and violation\_description tables. Because you’re appending the data (i.e., adding rows) to the end of the table instead of joining the data (i.e., adding columns to the table), you need to use a Union.

Drag the NYCParkingData source to Join 1, highlighting Union (not Join):



and you’ll see this:


4. Select Union 1 will show the results of appending NYCParkingData to the end of your dataset. You should have 10 “mismatching” fields:
* Plate\_ID
* Street\_Code1-Street\_Code2-Street\_Code3
* Law\_Section\_Sub\_Division
* Violation\_Legal\_Code
* Violation\_Description-1
* Summons\_Number-1
* Description
* Law\_SectionSub\_Division
* Summons\_Number-3
* Summons\_Number-2

If only these 10 fields are mismatched, then you have done the previous steps properly and you can move on to the next step. If there are other fields that are mismatched, double-check your column names (Tableau Prep treats “Day of Week” and “Day Of Week” as two different columns!) and data transformations to make sure you did everything correctly.

1. Add an Output step to the Union step and save it as a CSV file called “EditedNYCParkingData.csv”.
2. Open the file in Excel and answer the following questions by looking through the new data set:
* How many tickets were issues for driver’s registered in Pennsylvania (PA)?
(use a filter)
* How many tickets were issued for vehicles with a make of “Dodge”?
(use a filter)
* How many additional tickets were likely issued to Dodge vehicle but were missed in that previous count due to a misspelling?
(use a filter)
* What day of the week was the most common to get a ticket?
(create a chart by highlighting the appropriate column and inserting a “Recommended Chart”)