**Assignment #4: ETL using Tableau Prep**

|  |
| --- |
| **Submission Instructions**  **Due: Tuesday, 3/12/2019 at 9:30 am**   * Complete and submit ***TWO FILES*** below:   + The final version of your combined **data file** (“Class Performance.csv”)   + The Tableau Prep **flow file** you created to transform the data * If you do not follow the instructions, your assignment will be counted late.   + **Late Assignment policy**: All assignments will be assessed **a 50% penalty** (subtracted from that assignment’s score) for the first day (i.e. 24 hours) they are late***.*** |

You will be working with two datasets to complete this assignment. Both datasets contain similar information. The first dataset contains data about students’ math class performance. The second dataset contains data about students’ performance in a Portuguese class. Both datasets are available on the dataanalytics.temple.edu MySQL server. You will combine the two datasets and then transform the data using Tableau Prep.

Before you begin, review the data dictionary of the dataset on the last page of this document.

To start, create a connection to dataanalytics.temple.edu as a MySQL data source. Use your MySQL username and password. The data is located within the mxxgrades schema, as the math\_grades and portuguese\_grades tables.

Now complete the following steps:

1. Since the data sets have the same structure, start by combining them using a Union.
2. Separate all values so that each value has its own column (i.e., there isn’t more than one value per column). Refer to the data dictionary to see how many columns there should be.
3. Rename some of the fields according to the table below, making sure the data type is correct (refer to the Data Dictionary for the correct data types and then adjust in Tableau Prep accordingly):

|  |  |
| --- | --- |
| **Current Field Name:** | **Change to:** |
| G1 (Split1) | First Grade Raw Score |
| G2 (Split2) | Second Grade Raw Score |
| G3 (Split3) | Final Raw Score |
| Dalc | Weekday Alcohol Consumption |
| Walc | Weekend Alcohol Consumption |

1. Use Group and Replace to correct the data in the Mjob (Mother’s job) field. Pay attention to the five acceptable options listed in the data dictionary.
2. Use Group and Replace to correct the data in the Fjob (Father’s job) field. Pay attention to the five acceptable options listed in the data dictionary.
3. Create a calculated field to show grade percentage instead of raw score and name the field: Total Final Scores. (final raw score)

**HINT:**

To create a calculated field for Total Final Scores, first make sure that the Final Raw Score data type is set to numerical data. In the calculated field dialog, type the following for the formula:

([Final Raw Score]/20)\*100

1. Created a calculated field based on travel time using an IF statement to show the actual string values (refer to the Data Dictionary) based on the numeric values. Then name the field “Total Travel Time.”

**HINT:**

An IF statement is a conditional statement, that if true, will perform a function or display information. To create an IF statement for travel time, first look at the data dictionary for that field. The dictionary tells you that a value of 1 means a travel time of <15 minutes (less than 15 minutes), a value of 2 means 15 to 30 minutes, and so on. In the calculated field dialog, type the following for the formula:

IF [traveltime]= 1 THEN '<15 min.' ELSEIF [traveltime] = 2   
THEN '15-30 min.' ELSEIF [traveltime] = 3 THEN '30 min to 1 hour' ELSEIF [traveltime] = 4 THEN '>1 hour' END

The equation above is telling Tableau Prep that IF traveltime has a value of 1, then display ‘<15min.’ If the value is not 1, then check to see if the value is 2 (ELSEIF). The equation continues until it reaches END.

1. Create another calculated field with an IF statement for studytime to show the actual string values based on the numeric values. Name this field Total Study Time.
2. Export the data as “Class Performance.csv” and save your Tableau Prep file.

**Data Dictionary for math\_grade and portuguese\_grades tables**

|  |  |
| --- | --- |
| **Column** | **Description** |
| school | Student's school (binary: 'GP' or 'MS') |
| sex | Student's sex (binary: 'F' - female or 'M' - male) |
| age | Student's age (numeric: from 15 to 22) |
| address | Student's home address type (binary: 'U' - urban or 'R' - rural) |
| famsize | Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| Pstatus | Parent's cohabitation status (binary: 'T' - living together or 'A' - living apart) |
| Medu | Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education) |
| Fedu | Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education) |
| Mjob | Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| Fjob | Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') |
| reason | Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') |
| guardian | Student's guardian (nominal: 'mother', 'father' or 'other') |
| traveltime | Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| studytime | Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | Number of past class failures (numeric: n if 1<=n<3, else 4) |
| schoolsup | Extra educational support (binary: yes or no) |
| famsup | Family educational support (binary: yes or no) |
| paid | Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) |
| activities | Extra-curricular activities (binary: yes or no) |
| nursery | Attended nursery school (binary: yes or no) |
| higher | Wants to take higher education (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| romantic | With a romantic relationship (binary: yes or no) |
| famrel | Quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| freetime | Free time after school (numeric: from 1 - very low to 5 - very high) |
| goout | Going out with friends (numeric: from 1 - very low to 5 - very high) |
| Dalc | Workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| Walc | Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| health | Current health status (numeric: from 1 - very bad to 5 - very good) |
| absences | Number of school absences (numeric: from 0 to 93) |
| G1 | First period grade (numeric: from 0 to 20) |
| G2 | Second period grade (numeric: from 0 to 20) |
| G3 | Final grade (numeric: from 0 to 20, output target) |