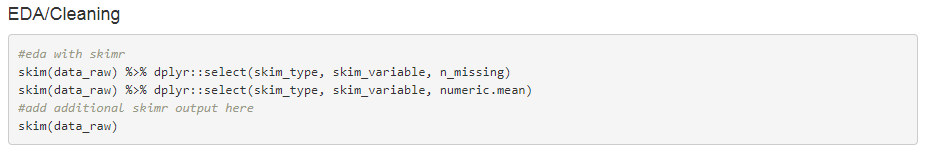
**Data Understanding**

Using the data provided via <https://smu.box.com/s/k9x192jxm39enjw2wx8ouw2kopx33l32>, the EDA task was completed and the model was created in R. The understanding of the tasks is to classify column “y” using features x0-x49 to minimize costs on a theoretical unknown dataset. If a false positive is generated it will cost $10 and if a false negative is generated it will cost $500, if true positive or negative is produced then the cost is $0.



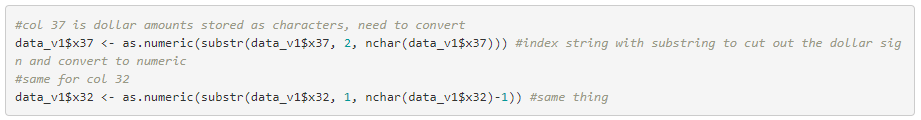
The skim function was called to provide a summary of the data frame. Any missing data is identified, variable means and variable histograms. Most attributes are missing a few, less than triple digits. Likely not significant compared to the 160K total rows. Also note that the average value of “y” is 0.401, indicating an estimate of 40% of the rows have y=1.



In addition, a review of rows that include “NA” was completed. There is roughly 1,500 rows that are removed that include “NA.” This makes up for 0.9% of the data.



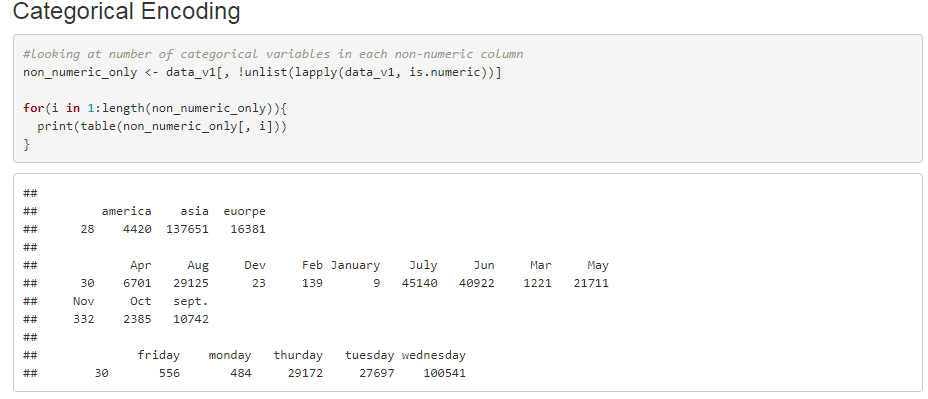
Attributes x37 and x32 are dollar figures and includes a “$” in the cell and the value is a character. This field needs to be converted into a numeric format. A function to index the string with a substring was generated to remove the dollar sign and convert to numeric for both attributes.



Another function to remove “NA” and another 70 rows are removed.



Encoding feature is ran due to a few other variables that require categorical change. First, take a glance at the number of categorical variables in each non-numeric column. Variable x24 appears to be a region or continent and includes “america,” “asia,” and “europe.” Respectively there 4,420 records for “america,” 137,651 records for “asia,” 16,381 records for “europe” and 28 records were blank. The same exercise was done for attributes x29 (month) and x30, “day of the week” excluding Saturday and Sunday but does comprise blank fields.



Numeric values were assigned to each region, with “america” = 1, “asia” = 2, “europe” = 3 and blank = 4. A “region\_numeric” data frame is created for later use in the model.



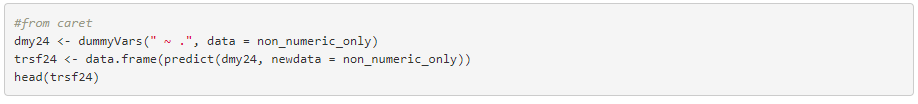
Monthly encoding is also performed to convert variable x29 (months) into numeric form. A “month\_numeric” data frame is created for later use in the model building.



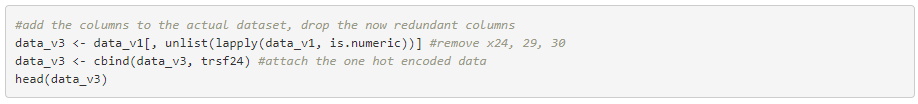
Finally the same process is completed for attribute x30 (day of week) and “day of the week\_numeric” data frame is created to use in the model building step. Each of these categorical changes were vetted using by backing into the data using a table function to produce the output before and after the conversion.



Dummy variables are created for each of the categorical variables. Four new variables for region are created, 13 new variables for month and 5 new variables for day of the week. There are 23 regional values that are blank, **X amount of month values???** and 12 day of the week values are blank.



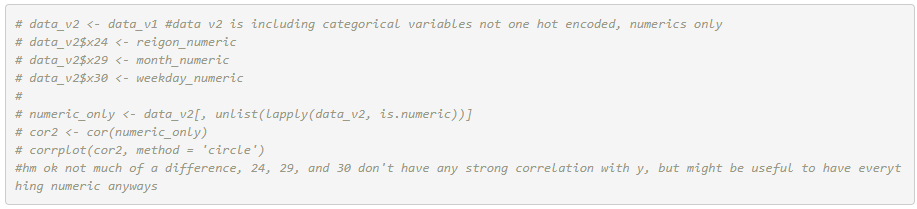
A new data frame is created that includes the new dummy variables. With a visual of the first 6 records in the new data frame.



Correlation Matrix and Analysis

Attributes x24, x29 and x30 are region, month and day of the week (weekday only) respectively. Each of these went through one-hot encoding which generated new dummy variables for binary values in each category. About 0.9% of the data missing with minimal value from imputing; therefore these rows values were removed from the data set.

The values were imputed into the new data frame including the new dummy variables in order to generate correlation analysis.



There are appear to be strong negative correlations between Wednesday and Tuesday, as well as asia and Europe. There does not appear to be strong correlation among any of the individual categorical variables and target y.

