## Function to join automotive data

def auto\_join(autos, makes):

import pandas as pd

key = ['make-id'] ## Define key column

## Return the joined dataframe

return pd.merge(autos, makes, on = key, how = 'left')

## Function to prepare the automotive data

def prep\_auto(df, col\_names):

import pandas as pd

import numpy as np

## Assign names to columns

df.columns = col\_names

## Drop unneeded columns

drop\_list = ['symboling', 'normalizedlosses', 'makeid']

df.drop(drop\_list, axis = 1, inplace = True)

## Remove rows with missing values

df = df[~pd.isnull(df).any(axis=1)]

## Add a log transformed column for price

df['lnprice'] = np.log(df['price'].as\_matrix())

## Remove duplicate rows

df.drop\_duplicates(inplace = True)

## Create a column with new levels for the number of cylinders

df['cylinders'] = ['four-or-less' if x in ['two', 'three', 'four'] else

('five-six' if x in ['five', 'six'] else

'eight-twelve') for x in df['cylinders']]

df = id\_outlier(df) # mark outliers

df = df[df.outlier == 0] # filter for outliers

df.drop('outlier', axis = 1, inplace = True)

return df

def id\_outlier(df):

## Create a vector of 0 of length equal to the number of rows

temp = [0] \* df.shape[0]

## test each outlier condition and mark with a 1 as required

for i, x in enumerate(df['enginesize']):

if (x > 190): temp[i] = 1

for i, x in enumerate(df['weight']):

if (x > 3500): temp[i] = 1

for i, x in enumerate(df['citympg']):

if (x > 40): temp[i] = 1

df['outlier'] = temp # append a column to the data frame

return df

def azureml\_main(autos, makes):

## join datasets

df = auto\_join(autos, makes)

## Define column names

col\_names = ['symboling', 'normalizedlosses', 'makeid', 'fueltype', 'aspiration', 'doors',

'body', 'drive', 'engineloc', 'wheelbase',

'length', 'width', 'height', 'weight', 'enginetype',

'cylinders', 'enginesize', 'fuelsystem', 'bore', 'stroke',

'compression', 'horsepower', 'rpm', 'citympg',

'highwaympg', 'price', 'make']

## Call function to prep auto data and return

return prep\_auto(df, col\_names)