Final Project_avanig_ndirks

December 13, 2015

Avani Goyal, Nathaniel Dirks : 12752 : Final Project Due: 12/13/2015

```
In [171]: import numpy as np
          import matplotlib.pyplot as plt
          import datetime as dt
          from operator import itemgetter
          import math
          %matplotlib inline
```

0.0.1 Load the RECS dataset into the memory.

It is loaded in two different variables to use it for two different purposes. 1. datanames: It stores RECS dataset along with the column names in tuple format 2. data1: It stores the data into a structured array format to be used for running iterations across all columns

```
In [172]: f= open('recs2009_public.csv','r')
          datanames = np.genfromtxt(f,delimiter=',', names=True,dtype=None)
```

```
In [173]: data1 = np.genfromtxt('recs2009_public.csv',delimiter=',', skip_header=1)
```

0.0.2 Preliminary analysis of dataset

The dataset is categorized for different regions such as 'midatlantic' (regional division - #2) and 'westsouthcentral' (regional division - #7)

```
In [174]: midatlantic = datanames[np.where(datanames['DIVISION']==2)]
          # print midatlantic[0]
          print midatlantic.shape
```

```
(1328,)
```

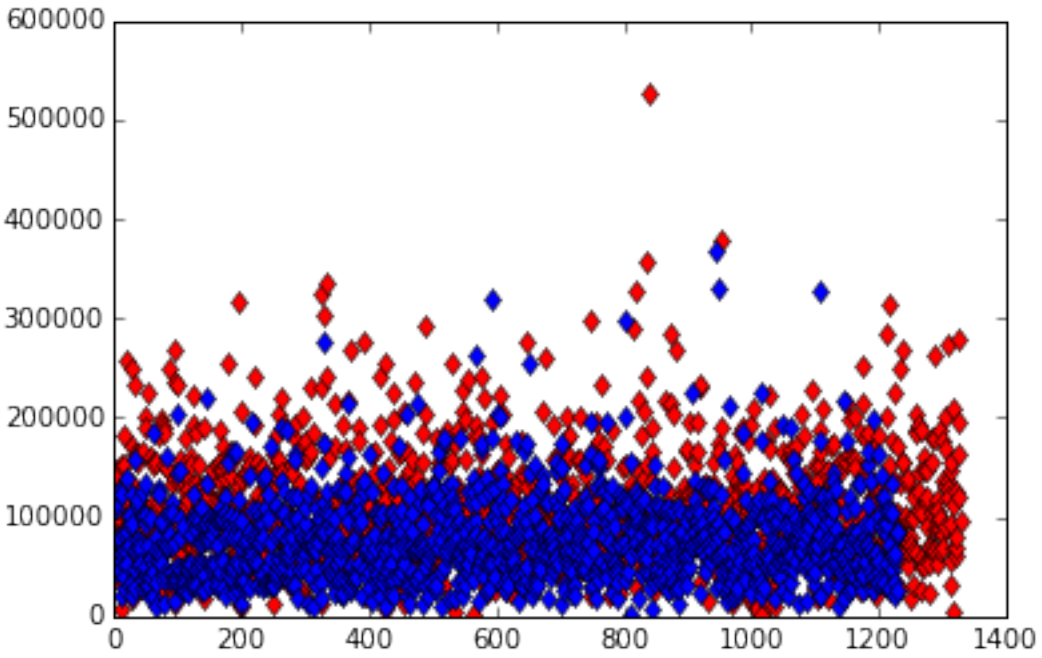
```
In [175]: wesouthcen = datanames[np.where(datanames['DIVISION']==7)]
          # wesouthcen[0]
          print wesouthcen.shape
```

```
(1230,)
```

'TOTALBTU' column represents the total energy consumption including electricity and other fuels like natural gas. Each regional dataset is plotted to observe the individual trends and to get a comparative picture.

```
In [176]: plt.plot(midatlantic['TOTALBTU'], 'rd')
          plt.plot(wesouthcen['TOTALBTU'], 'bd')
```

```
Out[176]: [<matplotlib.lines.Line2D at 0x1075d12d0>]
```

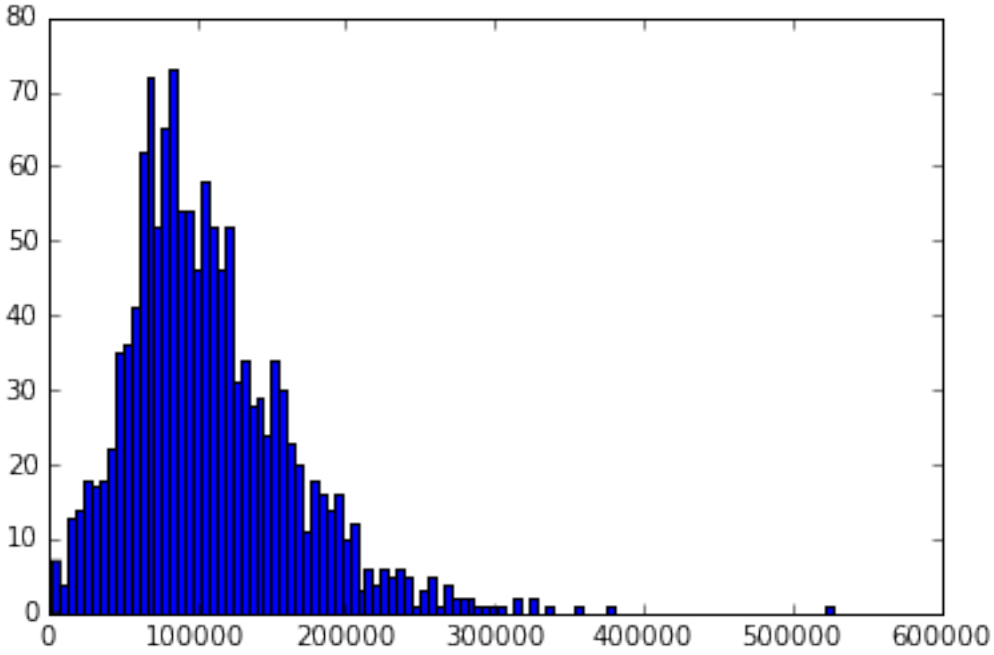


The individual trends are similar and show an almost linear horizontal line.

'MIDATLANTIC' region is selected for carrying out further analysis and build a regression model for predicting energy consumption values.

```
In [177]: plt.hist(midatlantic['TOTALBTU'],bins=100)

Out[177]: (array([  7.,   4.,  13.,  14.,  18.,  17.,  18.,  22.,  35.,  36.,  41.,
                   62.,  72.,  52.,  65.,  73.,  54.,  54.,  46.,  58.,  52.,  46.,
                   52.,  31.,  34.,  28.,  29.,  24.,  34.,  30.,  23.,  20.,  11.,
                   18.,  16.,  14.,  16.,  10.,  12.,   3.,   6.,   4.,   6.,   5.,
                    6.,   5.,   1.,   3.,   5.,   1.,   4.,   2.,   2.,   2.,   1.,
                    1.,   1.,   1.,   0.,   2.,   0.,   2.,   0.,   1.,   0.,   0.,
                    0.,   1.,   0.,   0.,   0.,   1.,   0.,   0.,   0.,   0.,   0.,
                    0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
                    0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
                    1.]),
           array([   3020.  ,    8263.35,   13506.7 ,   18750.05,   23993.4 ,
                    29236.75,   34480.1 ,   39723.45,   44966.8 ,   50210.15,
                    55453.5 ,   60696.85,   65940.2 ,   71183.55,   76426.9 ,
                    81670.25,   86913.6 ,   92156.95,   97400.3 ,  102643.65,
                   107887.  ,  113130.35,  118373.7 ,  123617.05,  128860.4 ,
                   134103.75,  139347.1 ,  144590.45,  149833.8 ,  155077.15,
                   160320.5 ,  165563.85,  170807.2 ,  176050.55,  181293.9 ,
                   186537.25,  191780.6 ,  197023.95,  202267.3 ,  207510.65,
                   212754.  ,  217997.35,  223240.7 ,  228484.05,  233727.4 ,
                   238970.75,  244214.1 ,  249457.45,  254700.8 ,  259944.15,
                   265187.5 ,  270430.85,  275674.2 ,  280917.55,  286160.9 ,
                   291404.25,  296647.6 ,  301890.95,  307134.3 ,  312377.65,
                   317621.  ,  322864.35,  328107.7 ,  333351.05,  338594.4 ,
                   343837.75,  349081.1 ,  354324.45,  359567.8 ,  364811.15,
```

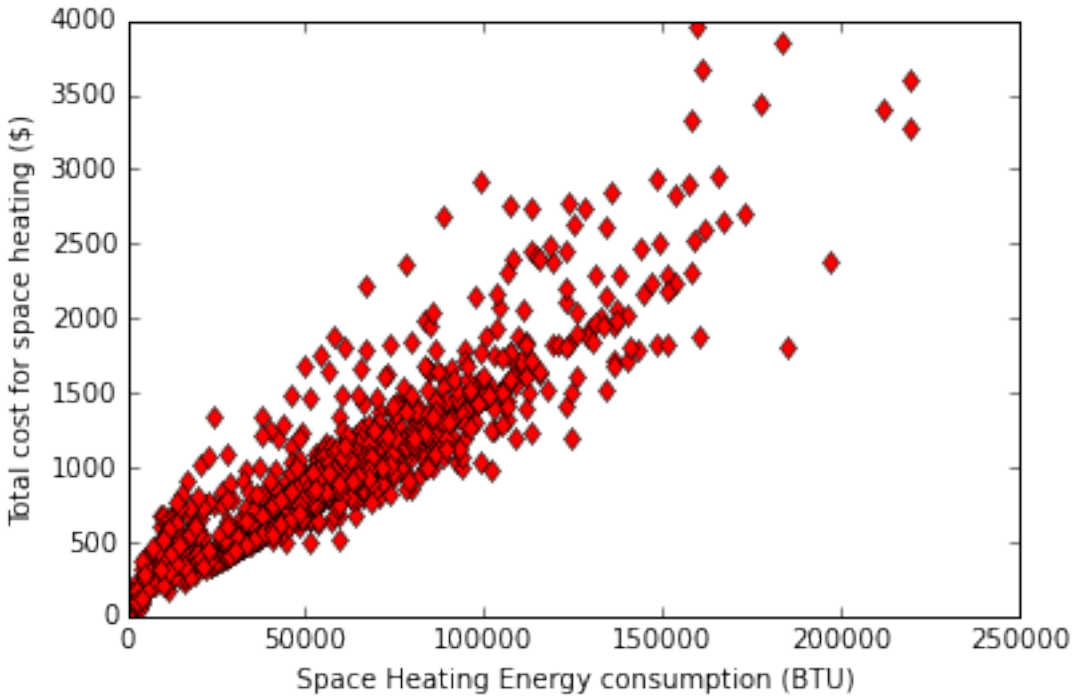
```
        370054.5 ,  375297.85,  380541.2 ,  385784.55,  391027.9 ,
        396271.25,  401514.6 ,  406757.95,  412001.3 ,  417244.65,
        422488.  ,  427731.35,  432974.7 ,  438218.05,  443461.4 ,
        448704.75,  453948.1 ,  459191.45,  464434.8 ,  469678.15,
        474921.5 ,  480164.85,  485408.2 ,  490651.55,  495894.9 ,
        501138.25,  506381.6 ,  511624.95,  516868.3 ,  522111.65,  527355.  ]),
 <a list of 100 Patch objects>)
```



Space heating energy consumption is analyzed against the dollar cost for space heating use to observe the correlation and check if it can be used for regression modeling.

```
In [208]: plt.plot(newdata['TOTALBTUSPH'],newdata['TOTALDOLSPH'], 'rd')
          plt.xlabel('Space Heating Energy consumption (BTU)')
          plt.ylabel('Total cost for space heating ($)')
```

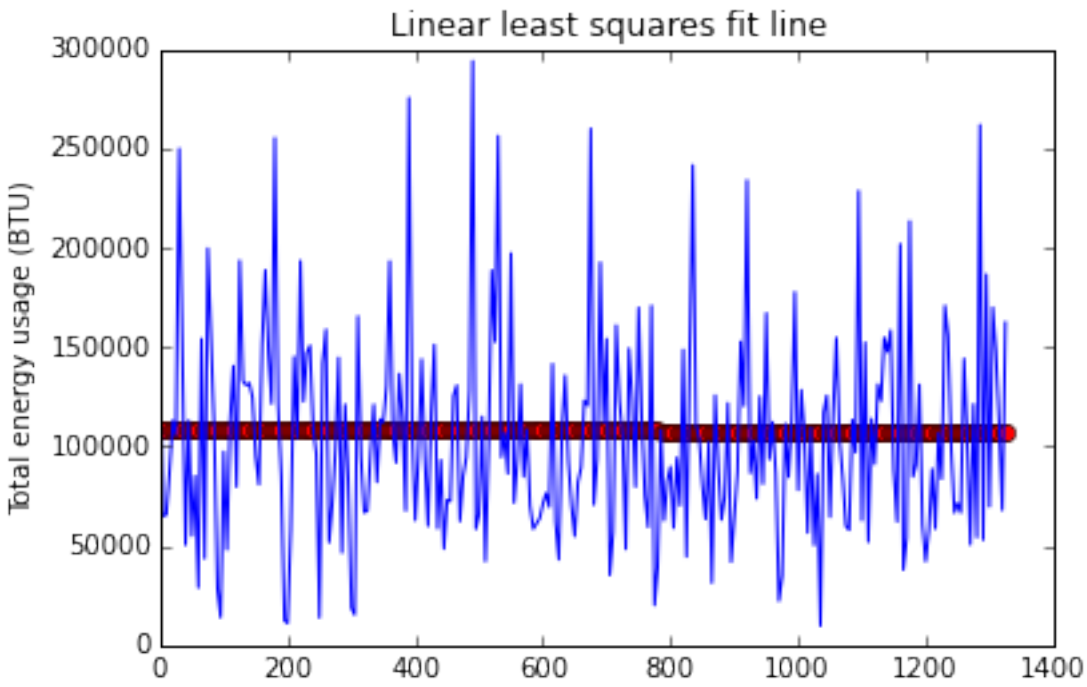
```
Out[208]: <matplotlib.text.Text at 0x10bab3e90>
```



0.1 Plotting a linear least squares fit line.

The line is observed to see the trendline of the randomly distributed data.

```
In [209]: xi = np.arange(0,1328)
          A = np.array([ xi, np.ones(1328)])
          # linearly generated sequence
          y = midatlantic['TOTALBTU']

          # obtaining the parameters
          w = np.linalg.lstsq(A.T,y)[0]

          xa = np.arange(0,1328,5)
          y = y[0:-1:5]

          # plotting the regression line
          line = w[0]*xa+w[1]
          plt.plot(xa,line,'ro',xa,y)
          plt.title('Linear least squares fit line')
          plt.ylabel('Total energy usage (BTU)')
          plt.show()

          print "Average value of energy consumption (BTU):"
          print np.average(y)
```




```
Average value of energy consumption (BTU):
104896.383459
```

The least square fit line is observed to be almost horizontal suggesting uniform distribution of the data across the mean value of 104,896 BTU.

0.1.1 Selection of highest correlated variables impacting total energy consumption.

Preliminarily, names and explanation of the variables are obtained by the 'public layout' file.

```
In [180]: names = np.genfromtxt('public_layout.csv', delimiter=',',skip_header=1,dtype=None,usecols=[1]
          print names
```

```
['Unique identifier for each respondent' 'Census Region' 'Census Division'
 'Reportable states and groups of states' 'Type of housing unit'
 'Final sample weight' '"Heating degree days in 2009'
 '"Cooling degree days in 2009' '"Heating degree days'
 '"Cooling degree days'
 'Building America Climate Region (collapsed for public file)'
 '"AIA Climate Zone'
 'Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area'
 'Housing unit classified as urban or rural by Census'
 '"Housing unit is owned' 'Housing unit part of condominium or cooperative'
 'Year housing unit was built' 'Year range when housing unit was built'
 'Year range when household moved in'
 'Converted 2-4 unit apartment building'
 'Converted 2-4 unit apartment building was originally a single-family house'
 'Converted 2-4 unit apartment building more like single family house or apartment building'
 'Number of floors in a 5+ unit apartment building'
```

'Number of apartment units in a 5+ unit apartment building'
'Major outside wall material' 'Major roofing material' 'Studio apartment'
'Number of floors in an apartment (Number of levels in housing unit that is an apartment)'
'Number of stories in a single-family home' 'Addition to a mobile home'
'Number of bedrooms' 'Number of full bathrooms' 'Number of half bathrooms'
'Number of rooms other than bedroom(s) and bathroom(s)'
'Total number of rooms in the housing unit' 'Basement in housing unit'
'Housing unit over a crawl space' 'Housing unit over a concrete slab'
'Finished basement' 'Number of finished rooms in the basement'
'Heating used in basement' 'All or partial basement heating'
'Portion of the basement which is heated' 'Cooling used in basement'
'All or partial basement cooling'
'Portion of the basement which is cooled'
'Portion of basement exclusively used by housing unit in apartment building with 2-4 units'
'Attic in housing unit' 'Finished attic'
'Number of finished rooms in the attic' 'Heating used in attic'
'All or partial attic heating' 'Portion of the attic which is heated'
'Cooling used in attic' 'All or partial attic cooling'
'Portion of the attic which is cooled'
'Portion of attic exclusively used by housing unit in apartment building with 2-4 units'
'Attached garage' 'Size of attached garage' 'Location of attached garage'
'Heating used in attached garage' 'Cooling used in attached garage'
'Detached garage or carport' 'Size of detached garage or carport'
'Outlet within 20 feet of vehicle parking' 'Imputation flag for KOWNRENT'
'Imputation flag for CONDCOOP' 'Imputation flag for YEARMADE'
'Imputation flag for YEARMADERANGE' 'Imputation flag for OCCUPYYRANGE'
'Imputation flag for CONVERSION' 'Imputation flag for ORIG1FAM'
'Imputation flag for LOOKLIKE' 'Imputation flag for NUMFLRS'
'Imputation flag for NUMAPTS' 'Imputation flag for WALLTYPE'
'Imputation flag for ROOFTYPE' 'Imputation flag for STUDIO'
'Imputation flag for NAPTFLRS' 'Imputation flag for STORIES'
'Imputation flag for TYPEHUQ4' 'Imputation flag for BEDROOMS'
'Imputation flag for NCOMBATH' 'Imputation flag for NHAFBATH'
'Imputation flag for OTHROOMS' 'Imputation flag for CELLAR'
'Imputation flag for CRAWL' 'Imputation flag for CONCRETE'
'Imputation flag for BASEFIN' 'Imputation flag for FINBASERMS'
'Imputation flag for BASEHEAT' 'Imputation flag for BASEHT2'
'Imputation flag for PCTBSTHT' 'Imputation flag for BASECOOL'
'Imputation flag for BASECL2' 'Imputation flag for PCTBSTCL'
'Imputation flag for BASEUSE' 'Imputation flag for ATTIC'
'Imputation flag for ATTICFIN' 'Imputation flag for FINATTRMS'
'Imputation flag for ATTCHEAT' 'Imputation flag for ATTCHT2'
'Imputation flag for PCTATTHT' 'Imputation flag for ATTCCOOL'
'Imputation flag for PCTATTCL' 'Imputation flag for ATTCCL2'
'Imputation flag for ATTICUSE' 'Imputation flag for PRKGPLC1'
'Imputation flag for SIZEOFGARAGE' 'Imputation flag for GARGLOC'
'Imputation flag for GARGHEAT' 'Imputation flag for GARGCOOL'
'Imputation flag for PRKGPLC2' 'Imputation flag for SIZEOFDETACH'
'Number of stoves (one appliance with cooktop and an oven)'
'Fuel used by most-used stove' 'Number of separate cooktops'
'Fuel used by most-used separate cooktop' 'Number of separate ovens'
'Fuel used by separate oven' 'Frequency of oven use' 'Self-cleaning oven'
'Continuous or manual cleaning cycle for most-used oven'
'Microwave oven used' 'Microwave usage' 'Microwave used for defrosting'

'Outdoor grill used' 'Fuel used by outdoor grill'
'Built-in indoor grill used' 'Fuel used by built-in indoor grill'
'Toaster used' 'Frequency hot meals are cooked' 'Most-used cooking fuel'
'Coffee maker used' 'Number of refrigerators used'
'Door arrangement of most-used refrigerator'
'Size of most-used refrigerator'
'Defrosting type of most-used refrigerator'
'Through-the-door ice and water on most-used refrigerator'
'Age of most-used refrigerator' 'Energy Star most-used refrigerator'
'Most-used refrigerator replaced by this household in the last 4 years'
'Assistance for replacing most-used refrigerator'
'Year of assistance for most-used refrigerator'
'Door arrangement of second most-used refrigerator'
'Size of second most-used refrigerator'
'Defrosting type of second most-used refrigerator'
'Number of months second most-used refrigerator used in 2009'
'Age of second most-used refrigerator'
'Energy Star second most-used refrigerator'
'Door arrangements of third most-used refrigerator'
'Size of third most-used refrigerator'
'Defrosting type of third most-used refrigerator'
'Number of months third most-used refrigerator used in 2009'
'Age of third most-used refrigerator'
'Energy Star third most-used refrigerator' 'Separate freezer used'
'Number of separate freezers used' 'Type of most-used freezer'
'Size of most-used freezer' 'Defrosting type for most-used freezer'
'Age of most-used freezer'
'Most-used freezer replaced by this household in the last 4 years'
'Assistance for replacing most-used freezer'
'Year of assistance for most-used freezer'
'Type of second most-used freezer' 'Size of second most-used freezer'
'Defrosting type for second most-used freezer'
'Age of second most-used freezer' 'Dishwasher used'
'Frequency dishwasher used' 'Age of dishwasher' 'Energy Star dishwasher'
'Dishwasher replaced by this household in the last 4 years'
'Assistance for replacing dishwasher' 'Year of assistance for dishwasher'
'Imputation flag for STOVEN' 'Imputation flag for STOVENFUEL'
'Imputation flag for STOVE' 'Imputation flag for STOVEFUEL'
'Imputation flag for OVEN' 'Imputation flag for OVENFUEL'
'Imputation flag for OVENUSE' 'Imputation flag for OVENCLN'
'Imputation flag for TYPECLN' 'Imputation flag for MICRO'
'Imputation flag for AMTMICRO' 'Imputation flag for DEFROST'
'Imputation flag for OUTGRILL' 'Imputation flag for OUTGRILL'
'Imputation flag for TOPGRILL' 'Imputation flag for STGRILA'
'Imputation flag for TOASTER' 'Imputation flag for NUMMEAL'
'Imputation flag for FUELFOOD' 'Imputation flag for COFFEE'
'Imputation flag for NUMFRIG' 'Imputation flag for TYPERFR1'
'Imputation flag for SIZRFRI1' 'Imputation flag for REFRIGT1'
'Imputation flag for ICE' 'Imputation flag for AGERFRI1'
'Imputation flag for TYPERFR2' 'Imputation flag for SIZRFRI2'
'Imputation flag for REFRIGT2' 'Imputation flag for MONRFRI2'
'Imputation flag for AGERFRI2' 'Imputation flag for TYPERFR3'
'Imputation flag for SIZRFRI3' 'Imputation flag for REFRIGT3'
'Imputation flag for MONRFRI3' 'Imputation flag for AGERFRI3'

'Imputation flag for SEPFREEZ' 'Imputation flag for NUMFREEZ'
'Imputation flag for UPRTFRZR' 'Imputation flag for SIZFREEZ'
'Imputation flag for FREEZER' 'Imputation flag for AGEFRZR'
'Imputation flag for UPRTFRZR2' 'Imputation flag for SIZFREEZ2'
'Imputation flag for FREEZER2' 'Imputation flag for AGEFRZR2'
'Imputation flag for DISHWASH' 'Imputation flag for DWASHUSE'
'Imputation flag for AGEDW' 'Clothes washer used in home'
'Top or front loading clothes washer used in home'
'Frequency clothes washer used'
'Water temperature used for clothes washer wash cycle'
'Water temperature used for clothes washer rinse cycle'
'Age of clothes washer' 'Energy Star clothes washer'
'Clothes washer replaced by this household in the last 4 years'
'Assistance for replacing clothes washer'
'Year of assistance for the clothes washer' 'Clothes dryer used in home'
'Fuel used by clothes dryer' 'Frequency clothes dryer used'
'Age of clothes dryer' 'Number of televisions used' 'Size of most-used TV'
'Display type of most-used TV'
'Cable box or satellite box connected to the most-used TV'
'DVR built into the cable box or satellite box connected to the most-used TV'
'Separate DVR connected to the most-used TV'
'Digital converter box connected to the most-used TV'
'Video game console connected to the most-used TV'
'Combo VCR/DVD connected to the most-used TV'
'VCR connected to the most-used TV'
'DVD player connected to the most-used TV'
'Home theater system connected to the most-used TV'
'Other set-top box connected to the most-used TV'
'Most-used TV usage on weekdays'
'Most-used TV weekday usage spent playing video games'
'Most-used TV usage on weekends'
'Most-used TV weekend usage spent playing video games'
'Size of second most-used TV' 'Display type of second most-used TV'
'Cable box or satellite box connected to the second most-used TV'
'DVR built into the cable box or satellite box connected to the second most-used TV'
'Separate DVR connected to the second most-used TV'
'Digital converter box connected to the second most-used TV'
'Video game console connected to the second most-used TV'
'Combo VCR/DVD connected to the second most-used TV'
'VCR connected to the second most-used TV'
'DVD player connected to the second most-used TV'
'Home theater system connected to the second most-used TV'
'Other set-top box connected to the second most-used TV'
'Second most-used TV usage on weekdays'
'Second most-used TV weekday usage spent playing video games'
'Second most-used TV usage on weekends'
'Second most-used TV weekend usage spent playing video games'
'Size of third most-used TV size' 'Display type of third most-used TV'
'Cable box or satellite box connected to the third most-used TV'
'DVR built into the cable box or satellite box connected to the third most-used TV'
'Separate DVR connected to the third most-used TV'
'Digital converter box connected to the third most-used TV'
'Video game console connected to the third most-used TV'
'Combo VCR/DVD connected to the third most-used TV'

'VCR connected to the third most-used TV'
'DVD player connected to the third most-used TV'
'Home theater system connected to the third most-used TV'
'Other set-top box connected to the third most-used TV'
'Third most-used TV usage on weekdays'
'Third most-used TV weekday usage spent playing video games'
'Third most-used TV usage on weekends'
'Third most-used TV weekend usage spent playing video games'
'Computer used at home' 'Number of computers used'
'Most-used computer - desktop or laptop'
'Monitor type of most-used computer' 'Daily usage of most-used computer'
'Turn off most-used computer when not in use'
'Sleep or standby mode for most-used computer when not in use'
'Second most-used computer - desktop or laptop'
'Monitor type of second most-used computer'
'Daily usage of second most-used computer'
'Turn off second most-used computer when not in use'
'Sleep or standby mode for second most-used computer when not in use'
'Third most-used computer - desktop or laptop'
'Monitor type of third most-used computer'
'Daily usage of third most-used computer'
'Turn off third most-used computer when not in use'
'Sleep or standby mode for third most-used computer when not in use'
'Internet access at home' 'Dial-up internet access'
'DSL or Fiber Optic internet access' 'Cable internet access'
'Satellite internet access' 'Wireless internet in home'
'Number of printers used' 'Separate fax machine used'
'Separate copy machine used' 'Well water pump used'
'Automotive block or engine heater or battery blanket used'
'Evaporative cooler used' 'Large heated aquarium used'
'Stereo equipment used' 'Cordless telephone used' 'Answering machine used'
'Number of rechargeable tools and appliances used'
'Charging patterns for rechargeable tools and appliances'
'Chargers for rechargeable tools and appliances left plugged into wall'
'Number of rechargeable electronic devices used'
'Charging patterns for rechargeable electronic devices'
'Chargers for rechargeable electronic devices left plugged into wall'
'Imputation flag for CWASHER' 'Imputation flag for TOPFRONT'
'Imputation flag for WASHLOAD' 'Imputation flag for WASHTEMP'
'Imputation flag for RNSETEMP' 'Imputation flag for AGECWASH'
'Imputation flag for DRYER' 'Imputation flag for DRYRFUEL'
'Imputation flag for DRYRUSE' 'Imputation flag for AGECDRYER'
'Imputation flag for TVCOLOR' 'Imputation flag for TVSIZE1'
'Imputation flag for TVTYPE1' 'Imputation flag for CABLESAT1'
'Imputation flag for COMBODVR1' 'Imputation flag for DVR1'
'Imputation flag for DIGITSTB1' 'Imputation flag for PLAYSTA1'
'Imputation flag for COMBOVCRDVD1' 'Imputation flag for VCR1'
'Imputation flag for DVD1' 'Imputation flag for TVAUDIOSYS1'
'Imputation flag for OTHERSTB1' 'Imputation flag for TVONWD1'
'Imputation flag for TVONWDWATCH1' 'Imputation flag for TVONWE1'
'Imputation flag for TVONWEWATCH1' 'Imputation flag for TVSIZE2'
'Imputation flag for TVTYPE2' 'Imputation flag for CABLESAT2'
'Imputation flag for COMBODVR2' 'Imputation flag for DVR2'
'Imputation flag for DIGITSTB2' 'Imputation flag for PLAYSTA2'

```
'Imputation flag for COMBOVCRDVD2' 'Imputation flag for VCR2'
'Imputation flag for DVD2' 'Imputation flag for TVAUDIOSYS2'
'Imputation flag for OTHERSTB2' 'Imputation flag for TVONWD2'
'Imputation flag for TVONWDWATCH2' 'Imputation flag for TVONWE2'
'Imputation flag for TVONWEWATCH2' 'Imputation flag for TVSIZE3'
'Imputation flag for TVTYPE3' 'Imputation flag for CABLESAT3'
'Imputation flag for COMBODVR3' 'Imputation flag for DVR3'
'Imputation flag for DIGITSTB3' 'Imputation flag for PLAYSTA3'
'Imputation flag for COMBOVCRDVD3' 'Imputation flag for VCR3'
'Imputation flag for DVD3' 'Imputation flag for TVAUDIOSYS3'
'Imputation flag for OTHERSTB3' 'Imputation flag for TVONWD3'
'Imputation flag for TVONWDWATCH3' 'Imputation flag for TVONWE3'
'Imputation flag for TVONWEWATCH3' 'Imputation flag for COMPUTER'
'Imputation flag for NUMPC' 'Imputation flag for PCTYPE1'
'Imputation flag for MONITOR1' 'Imputation flag for TIMEON1'
'Imputation flag for PCONOFF1' 'Imputation flag for PCSLEEP1'
'Imputation flag for PCTYPE2' 'Imputation flag for MONITOR2'
'Imputation flag for TIMEON2' 'Imputation flag for PCONOFF2'
'Imputation flag for PCSLEEP2' 'Imputation flag for PCTYPE3'
'Imputation flag for MONITOR3' 'Imputation flag for TIMEON3'
'Imputation flag for PCONOFF3' 'Imputation flag for PCSLEEP3'
'Imputation flag for INTERNET' 'Imputation flag for INDIALUP'
'Imputation flag for INDSL' 'Imputation flag for INCABLE'
'Imputation flag for INSATEL' 'Imputation flag for INWIRELESS'
'Imputation flag for PCPRINT' 'Imputation flag for FAX'
'Imputation flag for COPIER' 'Imputation flag for WELLPUMP'
'Imputation flag for DIPSTICK' 'Imputation flag for SWAMPCOL'
'Imputation flag for AQUARIUM' 'Imputation flag for STEREO'
'Imputation flag for NOCORD' 'Imputation flag for ANSMACH'
'Imputation flag for BATTOOLS' 'Imputation flag for BATCHRG'
'Imputation flag for CHRGPLGT' 'Imputation flag for ELECDEV'
'Imputation flag for ELECCHRG' 'Imputation flag for CHRGPLGE'
'Space heating equipment used' '"No space heating equipment'
'Unused space heating equipment type'
'Fuel for unused space heating equipment'
'Type of main space heating equpiment used' 'Main space heating fuel'
'Routine service or maintenance performed on main space heating equipment'
'Age of main space heating equipment'
'Main space heating equipment replaced by this household in the last 4 years'
'Assistance for replacing or maintaining main space heating equipment'
'Year of assistance for main space heating equipment'
'"Main space heating equipment heats other homes'
'Secondary space heating equipment used'
'Heat pump used for secondary space heating'
'Central warm-air furnace used for secondary space heating'
'Fuel used by warm-air furnace for secondary space heating'
'Hot water system used for secondary space heating'
'Fuel used by hot water system for secondary space heating'
'Built-in electric units used for secondary space heating'
'Pipeless furnace used for secondary space heating'
'Fuel used by pipeless furnace for secondary space heating'
'Built-in room heaters used for secondary space hearing'
'Fuel used by built-in electric units for secondary space heating'
'Heating stove used for secondary space heating'
```

'Fuel used by heating stove for secondary space heating'
'Portable electric heaters used for secondary space heating'
'Portable kerosene heaters used for secondary space heating'
'Fireplace used for secondary space heating'
'Fuel used by fireplace for secondary space heating'
'Flue on gas fireplace' 'Frequency gas fireplace used'
'Cooking stove used for secondary space heating'
'Fuel used by cooking stove for secondary space heating'
'Other equipment used for secondary space heating'
'Fuel used by other secondary space heating equipment'
'Portion of space heating provided by main space heating equipment (for homes with main and secondary
'Number of rooms heated' 'Thermostat(s) for heating equipment'
'Number of thermostats' 'Programmable main thermostat'
'Programmable thermostat lowers temperature at night'
'Programmable thermostat lowers temperature during the day'
'Temperature when someone is home during the day (winter)'
'Temperature when no on is home during the day (winter)'
'Temperature at night (winter)' 'Humidifier used'
'Number of months humidifier used in 2009' 'Imputation flag for HEATHOME'
'Imputation flag for DNTHEAT' 'Imputation flag for EQUIPNOHEAT'
'Imputation flag for FUELNOHEAT' 'Imputation flag for EQUIPM'
'Imputation flag for FUELHEAT' 'Imputation flag for MAINTHT'
'Imputation flag for EQUIPAGE' 'Imputation flag for HEATOTH'
'Imputation flag for FURNFUEL' 'Imputation flag for RADFUEL'
'Imputation flag for PIPEFUEL' 'Imputation flag for RMHTFUEL'
'Imputation flag for HSFUEL' 'Imputation flag for FPFUEL'
'Imputation flag for NGFPFLUE' 'Imputation flag for USENGFP'
'Imputation flag for RNGFUEL' 'Imputation flag for DIFFUEL'
'Imputation flag for EQMAMT' 'Imputation flag for HEATROOM'
'Imputation flag for THERMAIN' 'Imputation flag for NUMTHERM'
'Imputation flag for PROTHERM' 'Imputation flag for AUTOHEATNITE'
'Imputation flag for AUTOHEATDAY' 'Imputation flag for TEMPHOME'
'Imputation flag for TEMPGONE' 'Imputation flag for TEMPNITE'
'Imputation flag for MOISTURE' 'Imputation flag for USEMOISTURE'
'Number of tankless water heaters' 'Number of storage water heaters'
'Type of main water heater' 'Fuel used by main water heater'
'Main water heater is used by more than one housing unit'
'Main water heater size (if storage tank)' 'Main water heater age'
'Blanket around the main water heater (if storage tank)'
'Assistance for purchasing the water heater blanket'
'Year of the assistance for purchasing the water heater blanket'
'Type of secondary water heater' 'Fuel used by secondary water heater'
'Secondary water heater size (if storage tank)'
'Secondary water heater age' 'Imputation flag for NUMH2OHTRS'
'Imputation flag for NUMH2ONOTNK' 'Imputation flag for H2OTYPE1'
'Imputation flag for FUELH2O' 'Imputation flag for WHEATOTH'
'Imputation flag for WHEATSIZ' 'Imputation flag for WHEATAGE'
'Imputation flag for WHEATBKT' 'Imputation flag for H2OTYPE2'
'Imputation flag for FUELH2O2' 'Imputation flag for WHEATSIZ2'
'Imputation flag for WHEATAGE2' 'Air conditioning equipment used'
'"No air conditioning equipment'
'Type of unused air conditioning equipment'
'Type of air conditioning equipment used'
'Ducts for space heating and air conditioning'

```
'Central air conditioner is a heat pump'
'"Central air conditioner cools other homes'
'Routine service or maintenance performed on central air conditioner'
'Age of central air conditioner'
'Central air conditioner replaced by this household in the last 4 years'
'Assistance for maintaining or replacing central air conditioner'
'Year of assistance for central air conditioner' 'Number of rooms cooled'
'Frequency central air conditioner used in summer 2009'
'Thermostat for central air conditioner'
'Programmable thermostat for central air conditioner'
'Programmable thermostat adjusts temperature at night'
'Programmable thermostat adjusts temperature during the day'
'Temperature when someone is home during the day (summer)'
'Temperature when no on is home during the day (summer)'
'Temperature at night (summer)'
'Number of window/wall air conditioning units used'
'Age of most-used window/wall air conditioning unit'
'Energy Star most-used window/wall air conditioning unit'
'Most-used window/wall air conditioning unit replaced by this household in the last 4 years'
'Assistance for most-used window/wall air conditioning unit'
'Year of assistance for most-used window/wall air conditioning unit'
'Frequency most-used window/wall air conditioning unit used in summer 2009'
'Number of ceiling fans used'
'Frequency most-used ceiling fan used in summer 2009'
'Housing unit shaded from sun by large trees' 'Dehumidifier used'
'Number of months dehumidifier used in 2009' 'Imputation flag for AIRCOND'
'Imputation flag for DNTAC' 'Imputation flag for COOLTYPENOAC'
'Imputation flag for COOLTYPE' 'Imputation flag for DUCTS'
'Imputation flag for CENACHP' 'Imputation flag for ACOTHERS'
'Imputation flag for MAINTAC' 'Imputation flag for AGECENAC'
'Imputation flag for USECENAC' 'Imputation flag for ACROOMS'
'Imputation flag for THERMAINAC' 'Imputation flag for PROTHERMAC'
'Imputation flag for AUTOCOOLNITE' 'Imputation flag for AUTOCOOLDAY'
'Imputation flag for TEMPHOMEAC' 'Imputation flag for TEMPGONEAC'
'Imputation flag for TEMPNITEAC' 'Imputation flag for NUMBERAC'
'Imputation flag for WWACAGE' 'Imputation flag for USEWWAC'
'Imputation flag for NUMCFAN' 'Imputation flag for USECFAN'
'Imputation flag for TREESHAD' 'Imputation flag for NOTMOIST'
'Imputation flag for USENOTMOIST' 'High ceilings' 'Cathedral ceilings'
'Swimming pool' 'Heated swimming pool'
'Fuel used for heating swimming pool' 'Hot tub used'
'Fuel used for heating hot tub'
'Number of lights turned on 12 or more hours during a typical summer day'
'Number of energy-efficient bulbs for lights turned on 12 or more hours during a typical summer day'
'Number of lights turned on 4 to 12 hours during a typical summer day'
'Number of energy-efficient bulbs for lights turned on 4 to 12 hours during a typical summer day'
'Number of lights turned on 1 to 4 hours during a typical summer day'
'Number of energy-efficient bulbs for lights turned on 1 to 4 hours during a typical summer day'
'Number of outdoor lights left on all night'
'Number of energy-efficient bulbs for outdoor lights left on all night'
'Number of outdoor lights left on all night that use natural gas'
'Energy-efficient bulbs installed by this household'
'Assistance for energy-efficient light bulbs'
'Year of assistance for energy-efficient light bulbs'
```

```
'Sliding glass doors in heated areas'
'Number of sliding glass doors in heated areas'
'Number of windows in heated areas' 'Type of glass in most windows'
'Windows replaced by this household' 'Assistance for window replacement'
'Year of assistance for window replacement'
'Level of insulation (respondent reported)'
'Insulation added by this household' 'Year of added insulation'
'Assistance for added insulation'
'Year of assistance for added insulation'
'Is home too drafty in the winter? (respondent reported)'
'Caulking or weather stripping by this household'
'Year of caulking or weather stripping'
'Assistance for caulking or weather stripping'
'Year of assistance for caulking or weather stripping' 'Home energy audit'
'Year of home energy audit' 'Assistance for home energy audit'
'Year of assistance for home energy audit' 'Imputation flag for HIGHCEIL'
'Imputation flag for CATHCEIL' 'Imputation flag for SWIMPOOL'
'Imputation flag for POOL' 'Imputation flag for FUELPOOL'
'Imputation flag for RECBATH' 'Imputation flag for FUELTUB'
'Imputation flag for LGT12' 'Imputation flag for LGT4'
'Imputation flag for LGT1' 'Imputation flag for NOUTLGTNT'
'Imputation flag for NGASLIGHT' 'Imputation flag for SLDDRS'
'Imputation flag for DOOR1SUM' 'Imputation flag for WINDOWS'
'Imputation flag for TYPEGLASS' 'Imputation flag for NEWGLASS'
'Imputation flag for ADQINSUL' 'Imputation flag for INSTLINS'
'Imputation flag for AGEINS' 'Imputation flag for DRAFTY'
'Imputation flag for INSTLWS' 'Imputation flag for AGEWS'
'Imputation flag for AUDIT' 'Imputation flag for AGEAUD'
'Electricity is used in home' 'Natural gas is used in home'
'Propane is used in home' 'Fuel oil is used in home'
'Kerosene is used in home' 'Wood is used in home' 'Solar is used in home'
'Other fuel is used in home' 'Electricity used for space heating'
'Electricity used for secondary space heating'
'Electricity used for air conditioning'
'Electricity used for water heating' 'Electricity used for cooking'
'"Electricity used' 'Natural gas used for space heating'
'Natural gas used for secondary space heating'
'Natural gas used for water heating' 'Natural gas used for cooking'
'"Natural gas used' 'Propane used for space heating'
'Propane used for secondary space heating'
'Propane used for water heating' 'Propane used for cooking'
'"Propane used' 'Fuel oil used for space heating'
'Fuel oil used for secondary space heating'
'Fuel oil used for water heating' '"Fuel oil used'
'Kerosene used for space heating'
'Kerosene used for secondary space heating'
'Kerosene used for water heating' '"Kerosene used'
'Wood used for space heating' 'Wood used for secondary space heating'
'Wood used for water heating' '"Wood used' 'Solar used for space heating'
'Solar used for secondary space heating' 'Solar used for water heating'
'"Solar used' 'Other fuel used for space heating'
'Other fuel used for secondary space heating'
'Other fuel used for water heating' 'Other fuel used for cooking'
'Renewable on-site system used'
```

'Renewable on-site system connected to the grid'
'Who pays for electricity used for space heating'
'Who pays for electricity used for water heating'
'Who pays for electricity used for cooking'
'Who pays for electricity used for air conditioning'
'Who pays for electricity used for lighting and other appliances'
"Follow up for 'other' payment of electricity"
'Who pays for natural gas for space heating'
'Who pays for natural gas for water heating'
'Who pays for natural gas for cooking'
'Who pays for natural gas for other uses'
"Follow up for 'other' payment of natural gas" 'Who pays for fuel oil'
"Follow up for 'other' payment of fuel oil" 'Who pays for propane'
"Follow up for 'other' payment of propane" 'Propane delivered'
'Kerosene delivered to home' "Kerosene purchased 'cash and carry'"
"Number of kerosene 'cash and carry' purchases"
"Gallons per kerosene 'cash and carry' purchase" 'Wood logs used'
'Wood scraps used' 'Wood pellets used'
'"Type of wood used other than logs' 'Cords of wood used in 2009'
'Cords of wood used in 2009 (if more than 5)' 'Imputation flag for ONSITE'
'Imputation flag for ONSITEGRID' 'Imputation flag for PELHEAT'
'Imputation flag for PELHOTWA' 'Imputation flag for PELCOOK'
'Imputation flag for PELAC' 'Imputation flag for PELLIGHT'
'Imputation flag for OTHERWAYEL' 'Imputation flag for PGASHEAT'
'Imputation flag for PGASHTWA' 'Imputation flag for PUGCOOK'
'Imputation flag for PUGOTH' 'Imputation flag for OTHERWAYNG'
'Imputation flag for FOPAY' 'Imputation flag for OTHERWAYFO'
'Imputation flag for LPGPAY' 'Imputation flag for OTHERWAYLPG'
'Imputation flag for KERODEL' 'Imputation flag for KEROCASH'
'Imputation flag for NOCRCASH' 'Imputation flag for NKRGALNC'
'Imputation flag for WOODLOGS' 'Imputation flag for WDSCRAP'
'Imputation flag for WDPELLET' 'Imputation flag for WDOTHER'
'Imputation flag for WOODAMT' 'Imputation flag for NUMCORDS'
'Household fuel bills include fuel used for non-household purposes'
'Sex of householder' 'Employment status of householder'
'Householder lives with spouse or partner'
'Householder is Hispanic or Latino' "Householder's Race"
'Highest education completed by householder' 'Number of household members'
'Age of householder' 'Age category of second household member'
'Age category of third household member'
'Age category of fourth household member'
'Age category of fifth household member'
'Age category of sixth household member'
'Age category of seventh household member'
'Age category of eighth household member'
'Age category of ninth household member'
'Age category of tenth household member'
'Age category of eleventh household member'
'Age category of twelfth household member'
'Age category of thirteenth household member'
'Age category of fourteenth household member'
'Home-based business or service'
'Household member at home on typical week days'
'Household member(s) telecommutes or teleworks'

```
'Number of telecommuting days per month'
'Any activities that use an unusual amount of energy'
'Household members received employment income in 2009'
'Household members received retirement income in 2009'
'Household members received Supplemental Security income in 2009'
'Household members received welfare payments or cash assistance in 2009'
'Household members received investment income in 2009'
'Household members received other regular income in 2009'
'2009 gross household income'
'Household income at or below 100% of poverty line'
'Household income at or below 150% of poverty line'
'Housing unit in public housing authority' '"Lower rent due to Federal'
'Household receives food stamps or WIC assistance'
'Imputation flag for HHSEX' 'Imputation flag for HHAGE'
'Imputation flag for EMPLOYHH' 'Imputation flag for SPOUSE'
'Imputation flag for SDESCENT' 'Imputation flag for Householder_Race'
'Imputation flag for EDUCATION' 'Imputation flag for NHSLDMEM'
'Imputation flag for AGEHHMEMCAT2' 'Imputation flag for AGEHHMEMCAT3'
'Imputation flag for AGEHHMEMCAT4' 'Imputation flag for AGEHHMEMCAT5'
'Imputation flag for AGEHHMEMCAT6' 'Imputation flag for AGEHHMEMCAT7'
'Imputation flag for AGEHHMEMCAT8' 'Imputation flag for AGEHHMEMCAT9'
'Imputation flag for AGEHHMEMCAT10' 'Imputation flag for AGEHHMEMCAT11'
'Imputation flag for AGEHHMEMCAT12' 'Imputation flag for AGEHHMEMCAT13'
'Imputation flag for AGEHHMEMCAT14' 'Imputation flag for HBUSNESS'
'Imputation flag for ATHOME' 'Imputation flag for TELLWORK'
'Imputation flag for TELLDAYS' 'Imputation flag for OTHWORK'
'Imputation flag for WORKPAY' 'Imputation flag for RETIREPY'
'Imputation flag for SSINCOME' 'Imputation flag for CASHBEN'
'Imputation flag for INVESTMT' 'Imputation flag for RGLRPAY'
'Imputation flag for MONEYPY' 'Imputation flag for HUPROJ'
'Imputation flag for RENTHELP' 'Imputation flag for FOODASST'
'"Total square footage (includes all attached garages'
'"Total square footage (includes heated/cooled garages'
'Total heated square footage' 'Total unheated square footage'
'Total cooled square footage' 'Total uncooled square footage'
'Imputation flag for TOTSQFT' 'Imputation flag for TOTSQFT_EN'
'Imputation flag for TOTHSQFT' 'Imputation flag for TOTUSQFT'
'Imputation flag for TOTCSQFT' 'Imputation flag for TOTUCSQFT'
'"Total Site Electricity usage' '"Electricity usage for space heating'
'"Electricity usage for air-conditioning'
'"Electricity usage for water heating'
'"Electricity usage for refrigerators'
'"Electricity usage for other purposes (all end-uses except SPH'
'"Total Site Electricity usage' '"Electricity usage for space heating'
'"Electricity usage for air-conditioning'
'"Electricity usage for water heating'
'"Electricity usage for refrigerators'
'"Electricity usage for other purposes (all end-uses except SPH'
'"Total Electricity cost' '"Electricity cost for space heating'
'"Electricity cost for air-conditioning'
'"Electricity cost for water heating'
'"Electricity cost for refrigerators'
'"Electricity cost for other purposes (all end-uses except SPH'
'"Total Natural Gas usage' '"Natural Gas usage for space heating'
```

```
’"Natural Gas usage for water heating’
’"Natural Gas usage for other purposes (all end-uses except SPH and WTH)’
’"Total Natural Gas usage’ ’"Natural Gas usage for space heating’
’"Natural Gas usage for water heating’
’"Natural Gas usage for other purposes (all end-uses except SPH and WTH)’
’"Total Natural Gas cost’ ’"Natural Gas cost for space heating’
’"Natural Gas cost for water heating’
’"Natural Gas cost for other purposes (all end-uses except SPH and WTH)’
’"Total LPG/Propane usage’ ’"LPG/Propane usage for space heating’
’"LPG/Propane usage for water heating’
’"LPG/Propane usage for other purposes (all end-uses except SPH and WTH)’
’"Total LPG/LPG/Propane usage’ ’"LPG/Propane usage for space heating’
’"LPG/Propane usage for water heating’
’"LPG/Propane usage for other purposes (all end-uses except SPH and WTH)’
’"Total cost of LPG/Propane’ ’"Cost of LPG/Propane for space heating’
’"Cost of LPG/Propane for water heating’
’"Cost of LPG/Propane for other purposes (all end-uses except SPH and WTH)’
’"Total Fuel Oil usage’ ’"Fuel Oil usage for space heating’
’"Fuel Oil usage for water heating’
’"Fuel Oil usage for other purposes (all end-uses except SPH and WTH)’
’"Total Fuel Oil usage’ ’"Fuel Oil usage for space heating’
’"Fuel Oil usage for water heating’
’"Fuel Oil usage for other purposes (all end-uses except SPH and WTH)’
’"Total cost of Fuel Oil’ ’"Cost of Fuel Oil for space heating’
’"Cost of Fuel Oil for water heating’
’"Cost of Fuel Oil for other purposes (all end-uses except SPH and WTH)’
’"Total Kerosene usage’ ’"Kerosene usage for space heating’
’"Kerosene usage for water heating’
’"Kerosene usage for other purposes (all end-uses except SPH and WTH)’
’"Total Kerosene usage’ ’"Kerosene usage for space heating’
’"Kerosene usage for water heating’
’"Kerosene usage for other purposes (all end-uses except SPH and WTH)’
’"Total cost of Kerosene’ ’"Cost of Kerosene for space heating’
’"Cost of Kerosene for water heating’
’"Cost of Kerosene for other purposes (all end-uses except SPH and WTH)’
’"Total Wood usage’ ’"Total Wood usage’ ’"Total usage’
’"Total usage for space heating’ ’"Total usage for air conditioning’
’"Total usage for water heating’ ’"Total usage for refrigerators’
’"Total usage for appliances’ ’"Total cost’
’"Total cost for space heating’ ’"Total cost for air conditioning’
’"Total cost for water heating’ ’"Total cost for refrigerators’
’"Total cost for appliances’
’Electricity end uses included in Energy Supplier Survey billing data’
’Number of days covered by Energy Supplier Survey electricity billing data and used to calculate annual
’Whether annualized electricity consumption from Energy Supplier Survey billing data was scaled down’
’Natural gas end uses included in Energy Supplier Survey billing data’
’Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual
’Whether annualized natural gas consumption from Energy Supplier Survey billing data was scaled down’
’Number of days covered by Energy Supplier Survey LPG/propane billing data and used to calculate annual
’Whether annualized LPG/propane consumption from Energy Supplier Survey billing data was scaled down’
’Number of days covered by Energy Supplier Survey fuel oil billing data and used to calculate annual co
’Whether annualized fuel oil consumption from Energy Supplier Survey billing data was scaled down’
’Number of days covered by Energy Supplier Survey kerosene billing data and used to calculate annual co
’Whether annualized kerosene consumption from Energy Supplier Survey billing data was scaled down’]
```

Different variables are checked for their correlation value with the total energy consumption(TOTALBTU) based on manual understanding of the variables as shown below.

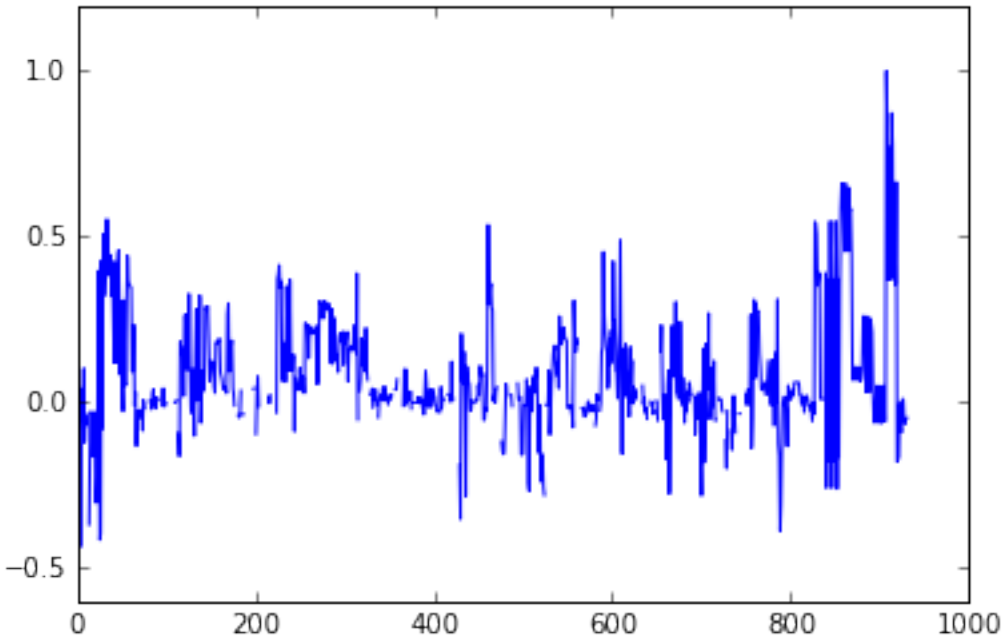
```
In [181]: np.corrcoef(midatlantic['WINDOWS'],midatlantic['TOTALBTU'])[1,0]

Out[181]: 0.48909065307980309

In [182]: np.corrcoef(midatlantic['TOTSQFT_EN'],midatlantic['TOTALBTU'])[1,0]

Out[182]: 0.53960276098516491

In [183]: np.corrcoef(midatlantic['TEMPHOME'],midatlantic['TOTALBTU'])[1,0]

Out[183]: 0.0038977065662747297

In [184]: np.corrcoef(midatlantic['NWEIGHT'],midatlantic['TOTALBTU'])[1,0]

Out[184]: -0.091091633291037893

In [185]: years = lambda d : ((dt.datetime.now()).year - d)
          yearsold = np.array(list(map(years, midatlantic['YEARMADE'])))
          midatlantic['YEARMADE']
          print yearsold

[93 95 47 ..., 67 50 75]

In [186]: np.corrcoef(midatlantic['YEARMADE'],midatlantic['TOTALBTU'])[1,0]

Out[186]: -0.027348298244598914

In [187]: np.corrcoef(midatlantic['TOTROOMS'],midatlantic['TOTALBTU'])[1,0]

Out[187]: 0.55248968425117162

In [188]: np.corrcoef(midatlantic['NHSLDMEM'],midatlantic['TOTALBTU'])[1,0]

Out[188]: 0.3098041035908895

In [189]: np.corrcoef(midatlantic['MONEYPY'],midatlantic['TOTALBTU'])[1,0]

Out[189]: 0.31068024918824672

In [190]: np.corrcoef(midatlantic['STORIES'],midatlantic['TOTALBTU'])[1,0]

Out[190]: 0.42837338142176584

In [191]: np.corrcoef(midatlantic['WASHTEMP'],midatlantic['TOTALBTU'])[1,0]

Out[191]: 0.34527156880018967
```

Result: The top factors based on the manual selection of variables are 'TOTSQFT_EN', 'TOTROOMS' and 'WINDOWS' with the correlation coefficient values ranging from 0.49 - 0.55.

This is further validated by running iteration using 'for' loop to obtain correlation coefficent values for all 931 variables.

```
In [192]: data1_ma = data1[(np.where(data1[:,2]==2))]

          def bestcorrelation(X):
              vector = np.zeros((len(X.T), 2))

              for i in range(len(X.T)):
                      vector[i,0] = int(i)
                      vector[i,1] = np.corrcoef(X[:,i],X[:,907])[1,0]
              return vector

          v = bestcorrelation(data1_ma)
          plt.plot(v[:,1])
          highcorr = v[(np.where(v[:,1]>=0.47))]
          print "Variable with correlation values greater than 0.53: "
          print highcorr
```

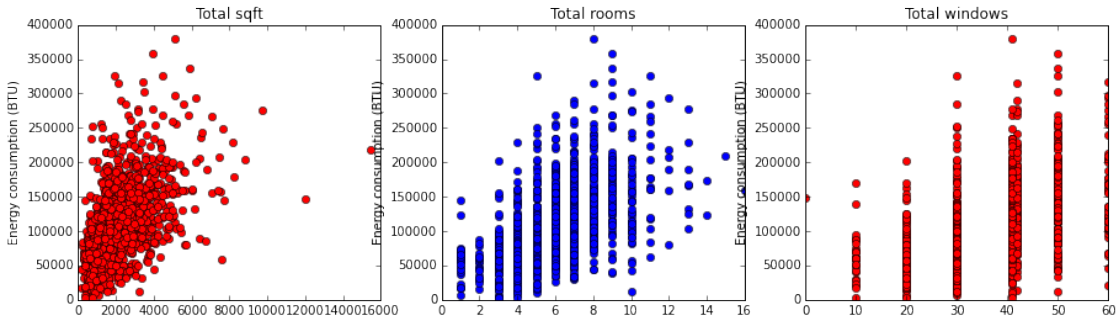
```
Variable with correlation values greater than 0.53:
[[  3.00000000e+01   5.08608923e-01]
 [  3.40000000e+01   5.52489684e-01]
 [  4.61000000e+02   5.35652360e-01]
 [  6.09000000e+02   4.89090653e-01]
 [  8.27000000e+02   5.44864210e-01]
 [  8.28000000e+02   5.39602761e-01]
 [  8.29000000e+02   5.07334111e-01]
 [  8.44000000e+02   5.46642960e-01]
 [  8.50000000e+02   5.46642823e-01]
 [  8.56000000e+02   5.83235640e-01]
 [  8.57000000e+02   6.60729926e-01]
 [  8.58000000e+02   5.87121765e-01]
 [  8.59000000e+02   6.02139506e-01]
 [  8.61000000e+02   6.60729395e-01]
 [  8.62000000e+02   5.87121153e-01]
 [  8.63000000e+02   6.02138516e-01]
 [  8.65000000e+02   6.47689633e-01]
 [  8.66000000e+02   5.75068831e-01]
 [  8.67000000e+02   5.84890620e-01]
 [  9.07000000e+02   1.00000000e+00]
 [  9.08000000e+02   8.63002777e-01]
 [  9.10000000e+02   7.69465065e-01]
 [  9.12000000e+02   6.83688101e-01]
 [  9.13000000e+02   8.72608607e-01]
 [  9.14000000e+02   7.78108186e-01]
 [  9.16000000e+02   5.90147697e-01]
 [  9.18000000e+02   6.64645304e-01]]
```



0.2 Multivariable regression modeling for midatlantic residential energy consumption

The top predictor variables are plotted against total the energy consumption values to visualize the trend.

```
In [193]: fig = plt.figure(1)
          fig.set_size_inches(15, 4)

          ax1 = fig.add_subplot(1,3,1)
          ax1.plot((data[:,0]),(data[:,3]),'ro')
          ax1.set_title("Total sqft")
          ax1.set_ylabel("Energy consumption (BTU)")

          ax2 = fig.add_subplot(1,3,2)
          ax2.plot((data[:,1]),(data[:,3]),'bo')
          ax2.set_title("Total rooms")
          ax2.set_ylabel("Energy consumption (BTU)")

          ax3 = fig.add_subplot(1,3,3)
          ax3.plot((data[:,2]),(data[:,3]),'ro')
          ax3.set_title("Total windows")

          ax3.set_ylabel("Energy consumption (BTU)")

          plt.show()
```



Base function for making designmatrix, beta_hat and R2 coefficents are defined for multi-variable regression modeling.

```
In [194]: def designmatrix(var1, var2, var3):
              designmatrix = np.vstack((var1, var2, var3))
              designmatrix = designmatrix.T
              return designmatrix

          def beta_hat(X,Y):
              dotp = np.dot(X.T,X)
              Ainv = np.linalg.inv(dotp)
              final = np.dot(Ainv,X.T)
              final = np.dot(final,Y)
              return final

          def R2(X,Y,beta_hat):
              m2 = Y-np.dot(X,beta_hat)
              m1 = m2.T
              y_avg =np.mean(Y)
              n2 = Y - y_avg
              n1 = n2.T
              R2_value = 1 - ((np.dot(m1,m2))/(np.dot(n1,n2)))
              return R2_value
```

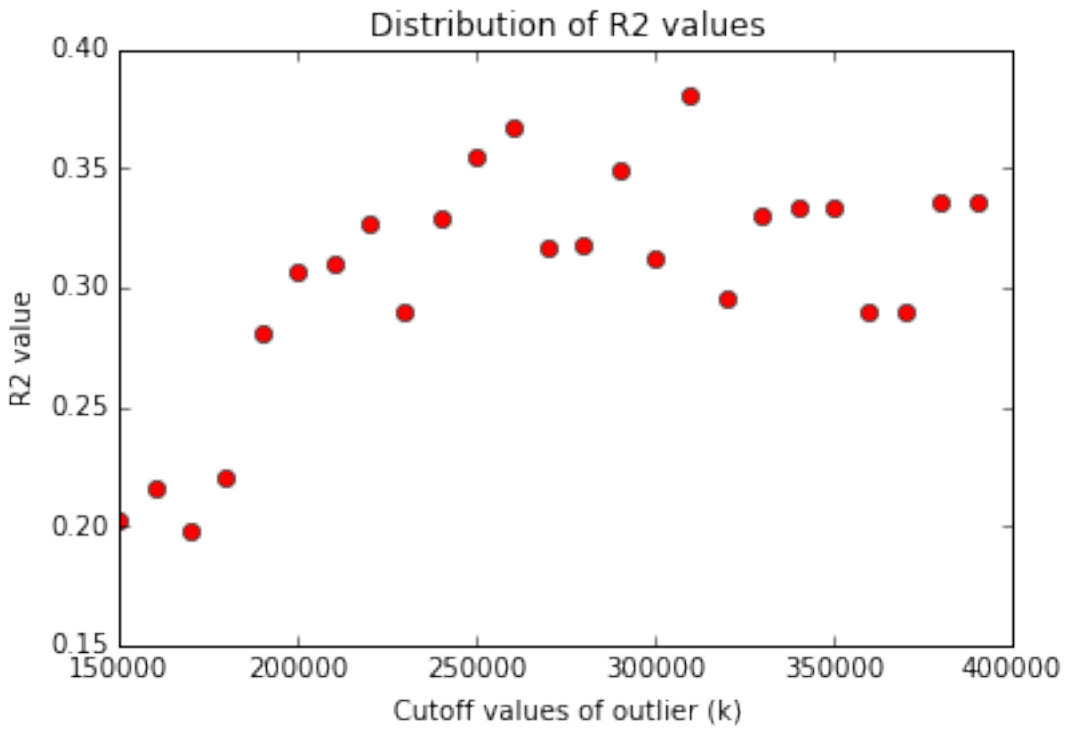
To remove the outliers, 'k' is defined as the cutoff above which the data will be trimmed. A 'for' loop is run below to optimize the 'k' value to obtain the maximum value of the R2 coefficient.

```
In [195]: R2_max = 0

          for k in range(150000,400000,10000):

              newdata = midatlantic[np.where(midatlantic['TOTALBTU']<k)]
              data = newdata['TOTSQFT_EN'],newdata['TOTROOMS'],newdata['WINDOWS'],newdata['TOTALBTU']
              data = np.transpose(data)

              data_sorted  = sorted(data, key=itemgetter(1))

              #Divide
              data = data[0:-1]
```

```
            data_train = data[::2]
            data_test = data[1::2]

            #Train dataset
            area_train = data_train[:,0]
            rooms_train = data_train[:,1]
            windows_train = data_train[:,2]
            btu_train = data_train[:,3]

            dmx1 = designmatrix(area_train,rooms_train,windows_train)
            beta_hat1 = beta_hat(dmx1,btu_train)

            #Test dataset
            area_test = data_test[:,0]
            rooms_test = data_test[:,1]
            windows_test = data_test[:,2]
            btu_test = data_test[:,3]

            dmx2 = designmatrix(area_test,rooms_test,windows_test)
            btu_pre = np.dot(dmx2,beta_hat1)

            R2_val = R2(dmx2,btu_test,beta_hat1)
            plt.plot(k,R2_val,'ro')
            plt.title('Distribution of R2 values')
            plt.xlabel('Cutoff values of outlier (k)')
            plt.ylabel('R2 value')

            if R2_max < R2_val:
                R2_max = R2_val
                k_max = k
            else:
                R2_max = R2_max
                k_max = k_max

        print "Maximum value of R2: ",R2_max
        print "At k value (k_max): ",k_max
        btu_test.shape
```

```
Maximum value of R2:  0.381068454214
At k value (k_max):  310000
```

```
Out[195]: (663,)
```



Using the results from above, the final dataset is created after removing the outliers having a value below k_max

```
In [196]: newdata = midatlantic[np.where(midatlantic['TOTALBTU']<k_max)]
          data = newdata['TOTSQFT_EN'],newdata['TOTROOMS'],newdata['WINDOWS'],newdata['TOTALBTU']
          data = np.transpose(data)
```

Split the final dataset into train and test data

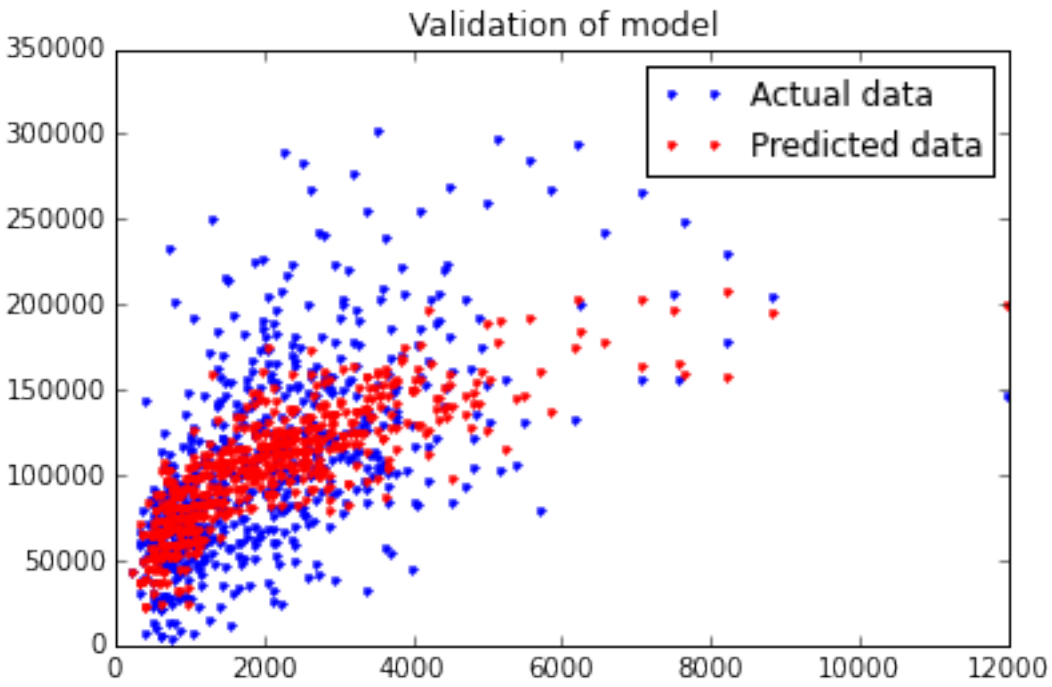
```
In [197]: # Data is sorted on number of total rooms
          data_sorted  = sorted(data, key=itemgetter(1))

          # Divide alternative values are taken henceforth for train and test dataset
          data_sorted = np.array(data_sorted[0:-1])
          data_train1 = np.array(data_sorted[::2])
          data_test1 = np.array(data_sorted[1::2])
          data_sorted

Out[197]: array([[  1188,      1,     41,  65043],
                 [   517,      1,     30,  36952],
                 [   313,      1,     10,  60251],
                 ...,
                 [  5047,     14,     60, 122480],
                 [  2043,     14,     50, 174227],
                 [  2748,     15,     60, 210048]])
```

0.3 Validation:

'Validation' function is created to build the model and make predictions for the energy consumption of test dataset.

It takes train dataset and test dataset as input and returns the R2 value and beta_matrix as output. It gives a plot to observe the comparison between actual and predicted values.

```
In [198]: def validation(data_train,data_test):

              #Train dataset
              btu_train = data_train[:,3]

              dmx1 = designmatrix(data_train[:,0],data_train[:,1],data_train[:,2])

              beta_hat1 = beta_hat(dmx1,btu_train)

              #Test dataset
              btu_test = data_test[:,3]

              dmx2 = designmatrix(data_test[:,0],data_test[:,1],data_test[:,2])

              btu_pre = np.dot(dmx2,beta_hat1)

              R2_val = R2(dmx2,btu_test,beta_hat1)

              print "R2 value is: ",R2_val

              plt.plot(data_test[:,0],btu_test,'.b')
              plt.plot(data_test[:,0],btu_pre,'.r')
              plt.legend(['Actual data','Predicted data'])
              plt.title('Validation of model')
              print "Beta matrix:",beta_hat1
              return (beta_hat1, R2_val)


          beta1, R2_1 = validation(data_train1,data_test1)

R2 value is:  0.400553308137
Beta matrix: [    6.40646137  6271.21455636  1476.27906611]
```



Mean of one variable is compared for both test and train dataset to check for significant difference between them.

```
In [199]: print np.mean(data_test[:,0])
          print np.mean(data_train[:,0])

2053.98039216
2096.43137255

In [200]: print np.mean(data_test[:,1])
          print np.mean(data_train[:,1])

5.84464555053
5.8838612368
```

0.4 Cross-validation:

The data has been split into three equal parts by selecting every third value for a dataset starting at different points.

```
In [201]: print data_sorted
          first = np.array(data_sorted[::3])
          second = np.array(data_sorted[1::3])
          third = np.array(data_sorted[2::3])

          print "First dataset[0]:",first[0]
          print "Second dataset[0]:",second[0]
          print "Third dataset[0]:",third[0]
```

```
[[  1188      1     41  65043]
 [   517      1     30  36952]
 [   313      1     10  60251]
 ...,
 [  5047     14     60 122480]
 [  2043     14     50 174227]
 [  2748     15     60 210048]]
First dataset[0]: [ 1188     1    41 65043]
Second dataset[0]: [  517     1    30 36952]
Third dataset[0]: [  313     1    10 60251]
```

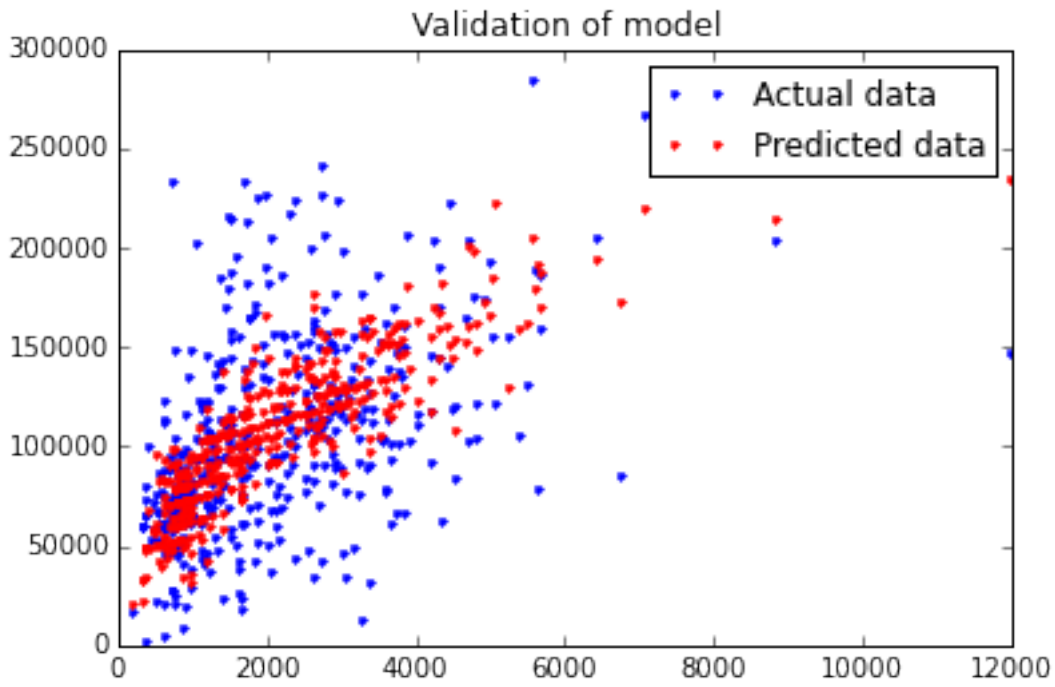
Three pairs of train and test datasets are created for cross validation purpose using the three datasets.

```
In [202]: data_train2 = np.vstack((first,second))
          data_test2 = np.array(third)
          print "Second split of datasets"
          print data_train2.shape
          print data_test2.shape

          data_train3 = np.vstack((first,third))
          data_test3 = np.array(second)
          print "Third split of datasets"
          print data_train3.shape
          print data_test3.shape

          data_train4 = np.vstack((third,second))
          data_test4 = np.array(first)
          print "Fourth split of datasets"
          print data_train4.shape
          print data_test4.shape
```

```
Second split of datasets
(880, 4)
(439, 4)
Third split of datasets
(879, 4)
(440, 4)
Fourth split of datasets
(879, 4)
(440, 4)
```

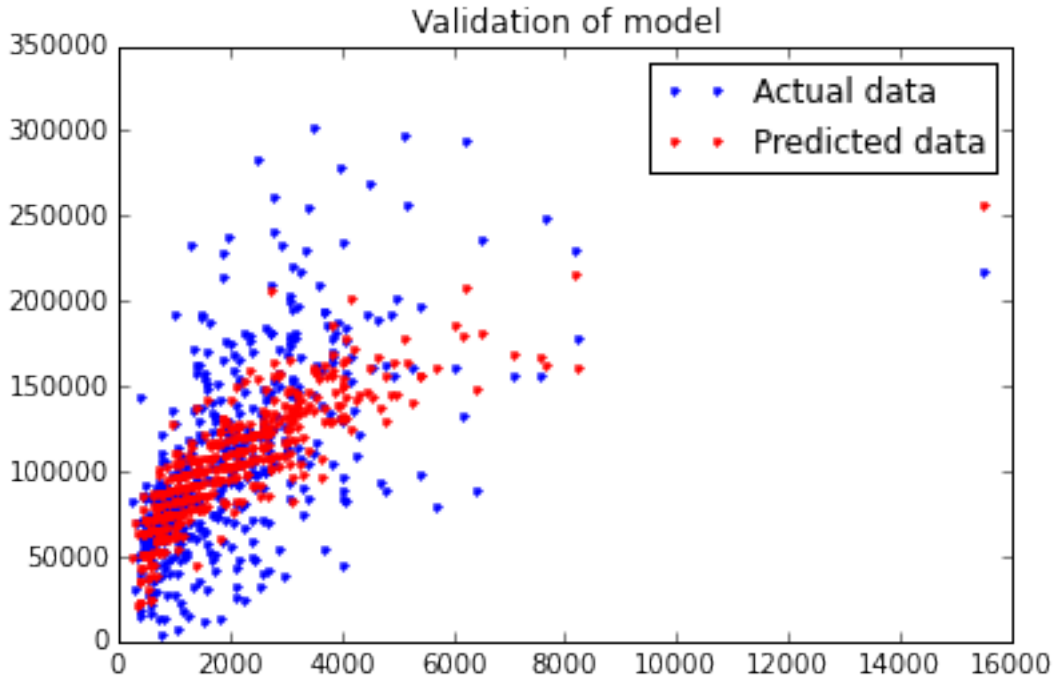
```
In [203]: beta2, R2_2 = validation(data_train2,data_test2)
```

```
R2 value is:  0.317785760003
Beta matrix: [    9.83361476  7469.64165002  1149.43064098]
```

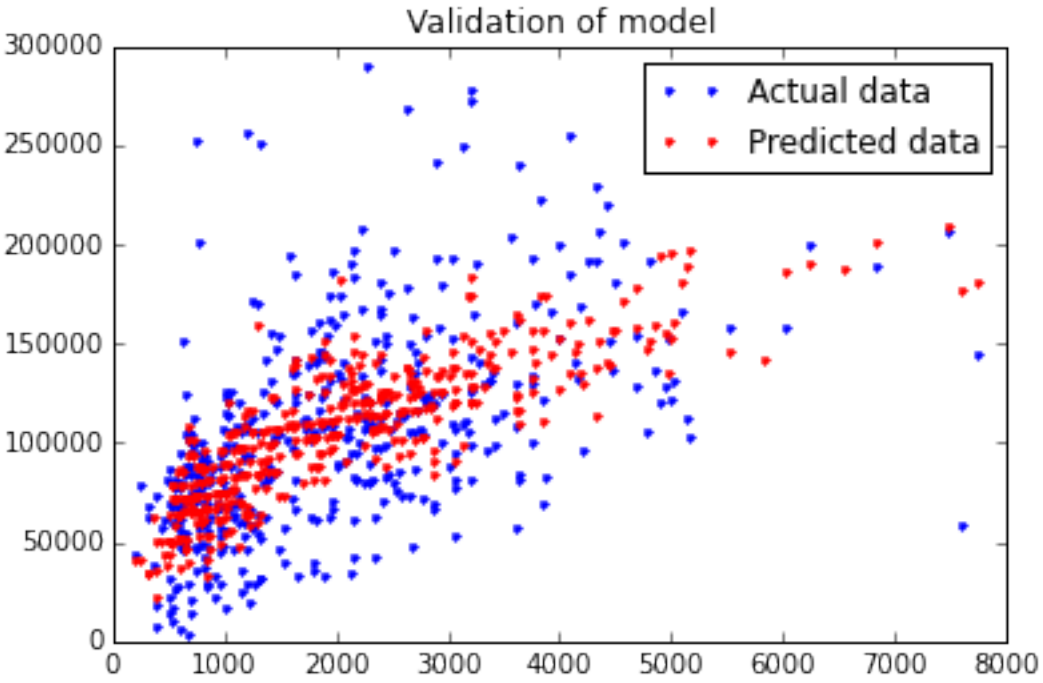


In [204]: beta3, R2_3 = validation(data_train3,data_test3)

```
R2 value is:  0.398525487668
Beta matrix: [  6.75388342e+00   7.34240501e+03   1.29650931e+03]
```



```
In [205]: beta4, R2_4 = validation(data_train4,data_test4)

R2 value is:  0.353451435755
Beta matrix: [    7.62732665  7409.65065618  1264.0514207 ]
```



Final Result: Mean values of R2 and Beta_hat matrices

```
In [206]: l = [R2_1,R2_2,R2_3,R2_4]
          R2_avg = np.mean(l)
          print "Mean R2 value: ",R2_avg

          beta_avg = np.mean([beta1,beta2,beta3,beta4],axis=0)
          print "Mean Beta_hat matrix: ",beta_avg

Mean R2 value:  0.367578997891
Mean Beta_hat matrix:  [    7.65532155  7123.22796853  1296.56760891]
```

Calculate uncertainties using 95% confidence intervals corresponding to t-distribution This is calculated using the first train dataset created and the average beta_hat matrix.

```
In [207]: # calculating error matrix: (Y-XB)
          btu_test = data_test1[:,3]
          dmx2 = designmatrix(data_test1[:,0],data_test1[:,1],data_test1[:,2])

          error = btu_test - np.dot(dmx2,beta_avg)
          # defining N for the number of data points in the test dataset
          N = error.size
          # defining the number of co-efficients in the beta_hat matrix
          p = beta_avg.size
```

```
        X = dmx2
        print "N=",N
        print "p=",p

        #squaring of error matrix is calculated by multiplying by its transpose
        errormatrix = (np.dot(error,error.T))/(N-p-1)
        # print "Standard mean error:",errormatrix

        s_var = errormatrix*(np.linalg.inv(np.dot(X.T,X)))
        # print s_var

        import math
        sqrt = lambda d: (math.sqrt(d))
        s_dev = map(sqrt,np.diag(s_var))
        # s_dev

        from scipy.stats import t
        T_val = t.isf((1-0.95)/2,(N-p-1))

        max_val = beta_avg + np.dot(T_val,s_dev)
        min_val = beta_avg - np.dot(T_val,s_dev)
        print "Base value: "+str(np.round(beta_avg, decimals=1))
        print "Maximum value: "+str(np.round(max_val, decimals=1))
        print "Minimum value: "+str(np.round(min_val, decimals=1))
```

```
N= 659
p= 3
Base value: [    7.7  7123.2  1296.6]
Maximum value: [   10.7  9344.1  1632.4]
Minimum value: [  4.70000000e+00   4.90230000e+03   9.60800000e+02]
```