MSDS 6372 Project 1 Report

Volodymyr Orlov, Jaime Villanueva, Jason Lin

Introduction

The United States energy consumption is one of the biggest in the world with residential sectors representing approximately 40 percent[1] of it. Here we explore data collected via *Residential Energy Consumption Survey*[2] conducted by the *US Energy Information Administration* in 2009. Using this data we perfom regression analysis of the data and model future average yearly energy consumption, in US dollars, of a household from various household's features and consumption habit of its tenants. In addition to that we also attempted to answer following questions:

- Q1
- Q2

Data Description

(Why we've chosen 2009 when we could take 2015?!!. Lets either somehow support it or use latest data).

The *US Residential Energy Consumption Survey* is a comprehensive data collection program that was conducted by the *US Energy Information Administration*[3].

The data file can be obtained from the www.eia.gov website and is formatted as a CSV file. The file contains responses from 12083 randomly selected sample of the population of 113.6 million US households. Response is coded using 940 variables, which can be broken down into following categories:

- 359 imputation flags.
- 85 measurements of various energy consumption metrics.
- 58 various quantitative measurements, collected from a household or its tenants.
- 438 categorical features of a household.

Exploratory Analysis

Fo our analisys and models we considered only 58 quantitative and 438 categorical features of a household and dropped 359 imputation flags as well as 85 measurements of energy consumption metrics. For our first model we've been primaraly interested in predicting total energy bill per household that has been recorded as a separate variable, in whole US dollars.

To better understand relationship between total yearly cost and household's features we looked at correlation matrix of all quantitative plus our dependent variable. The plot is depicted in Figure 1

By scanning through correlation matrix we've identified several good candidates for strong predictor variables: total square footage, total number of rooms, number of rooms heated, age of household, number of ceiling fans used, number of televisions used, heating degree days in 2009 above 65F, cooling degree days in 2009 below 65F. From **??** you can see there is a clear linear relashionship between some of these variables and a response we are trying to predict.

In addition to that we've used our intuition to select following variables for further analysis:

[1] https://www.nap.edu/catalog/13360/effective-tracking-of-building-energy-use-improving-the-commercial-buildings
[2] https://www.eia.gov/consumption/residential/about.php
[3] https://en.wikipedia.org/wiki/Energy_Information_Administration

Figures

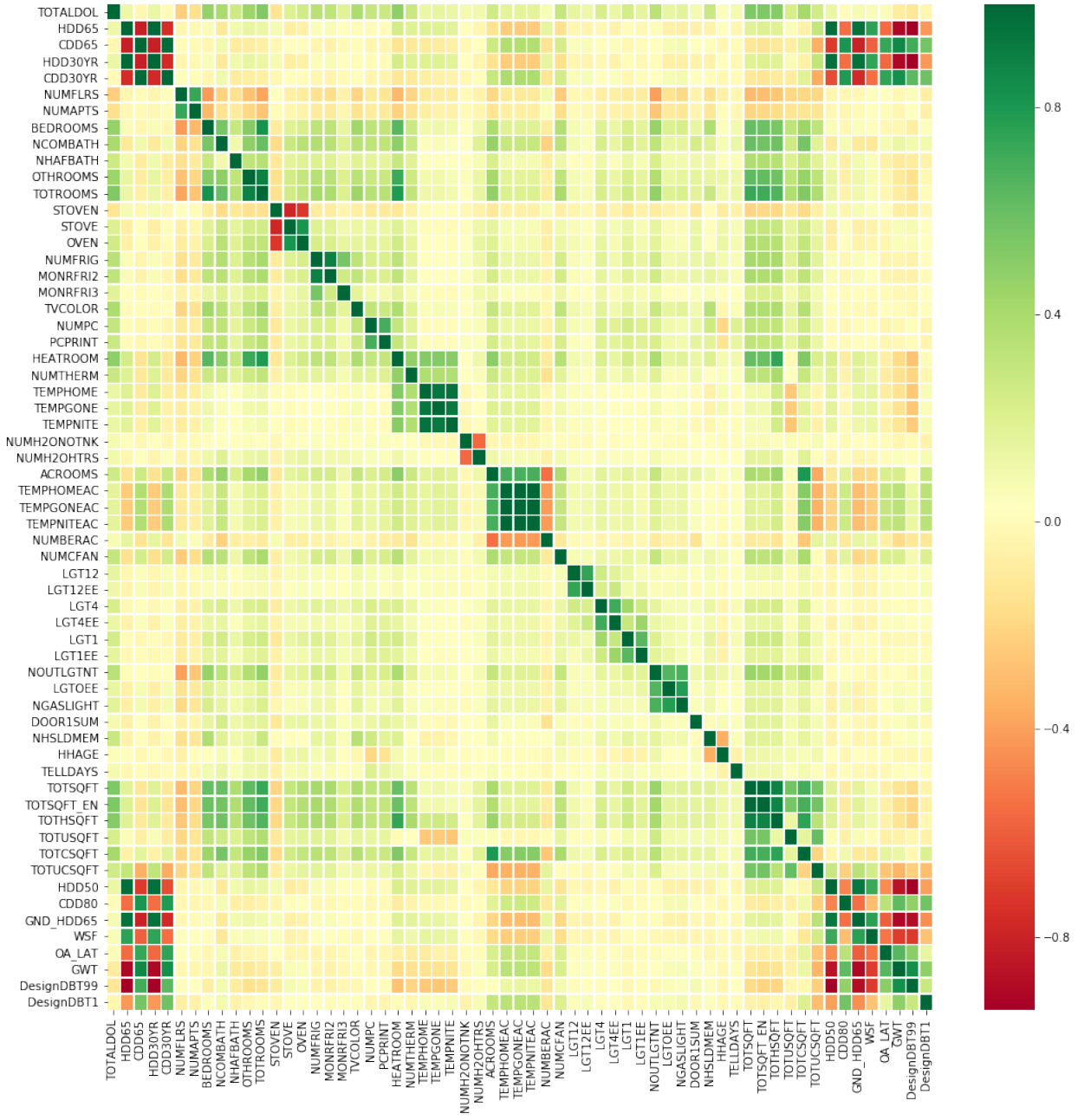

Figure 1: Correlation matrix of all quantitative household measurements, plus dependent variable.

Objective 1