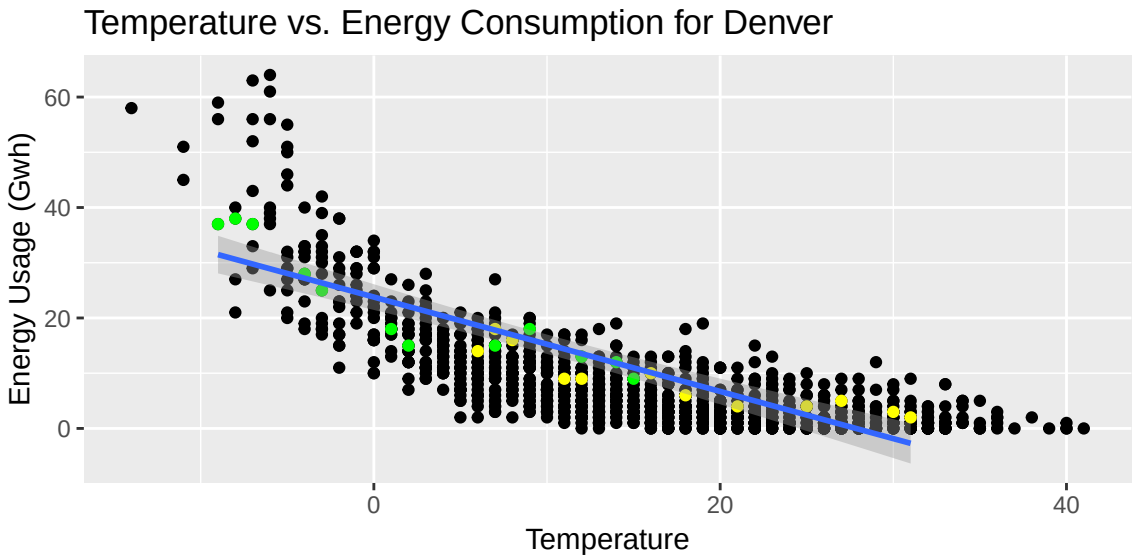
Final Report

Michael Ramsey

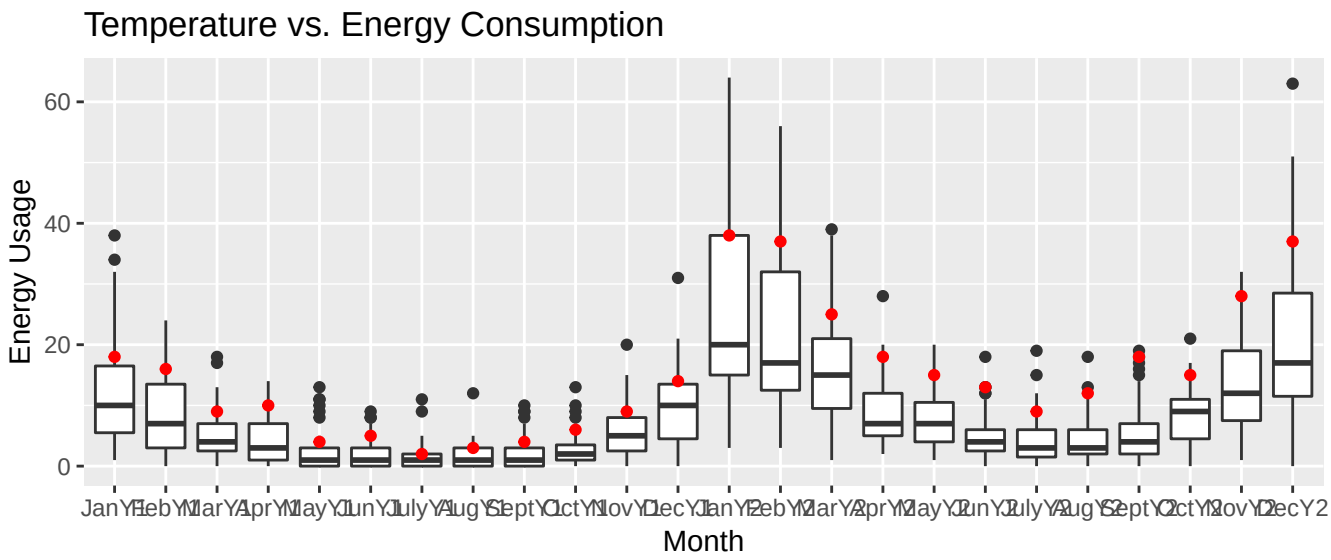
May 3, 2018

Exploratory Analysis

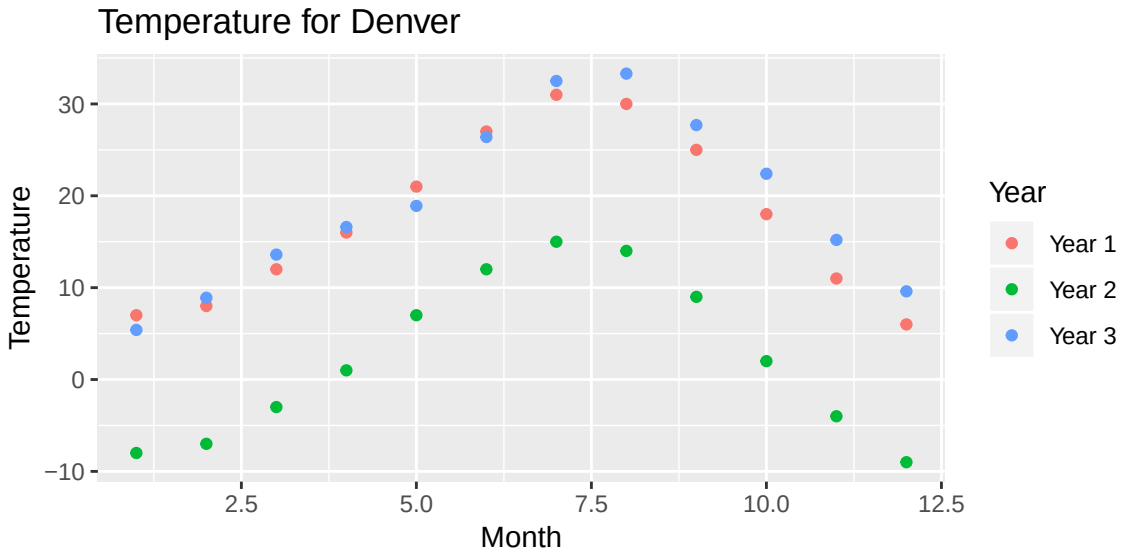
It is important for a data scientist to first explore the data when given a new problem. This will give the modeler intuition for applying the correct methods. I begin by plotting the two years of temperature and energy data that we have. I initially notice that the temperature for the second year of data is much lower than the first year. I highlight the specific values for Denver.



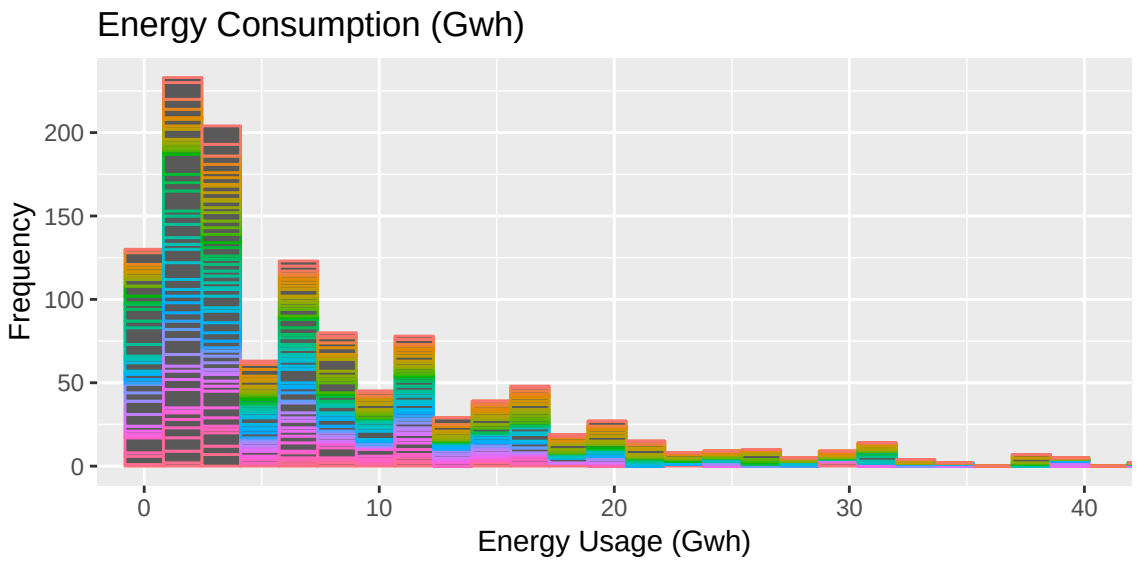
The yellow dots correspond to the data for the first year of Denver and the green dots correspond to the second year. We can see from the plot above that a normal linear model will likely be insufficient for predicting the temperature for a third year. At this point, I considered pooling some of the other city data with Denver. Consider the following box-plot of the energy usage by month for all of the cities. The listed temperature for Denver for all 24 months is listed by a red dot.



We see that the energy usage for Denver is well above the median city energy consumption for each month. Due to this, a pooled linear model will be unable to describe the energy consumption for Denver. Additionally, I searched through the data to find other cities with "similar" data points to Denver. If there are cities that have similar temperature and energy usage values to Denver, I could consider pooling that data. The "closest" cities to Denver are Washington DC and Richmond. However, even the data from these two cities did not closely resemble the data for Denver. Therefore I decided to not pool any of the data. Next I plot the temperature values by month for Denver.



At this point I was concerned about the big difference in temperatures for the two years of data for Denver. Since the temperature values for year 3 were much more similar to the values in year 1, I was worried that the data for year 2 would distort the predictions. However, I decided to keep them in the model. Next, I plot a histogram of the Energy Usage in Gwh.



Clearly the data is not normally distributed. I also plotted a histogram for the offset log-energy. This did make the data look more normally distributed, however not enough to be acceptable. Because our data is not normally distributed, any sort of linear model willl likely be ineffective at predicting the Denver temperature. I will have to consider a model that does not rely on normally distributed data. Specifically, I will consider poisson models.

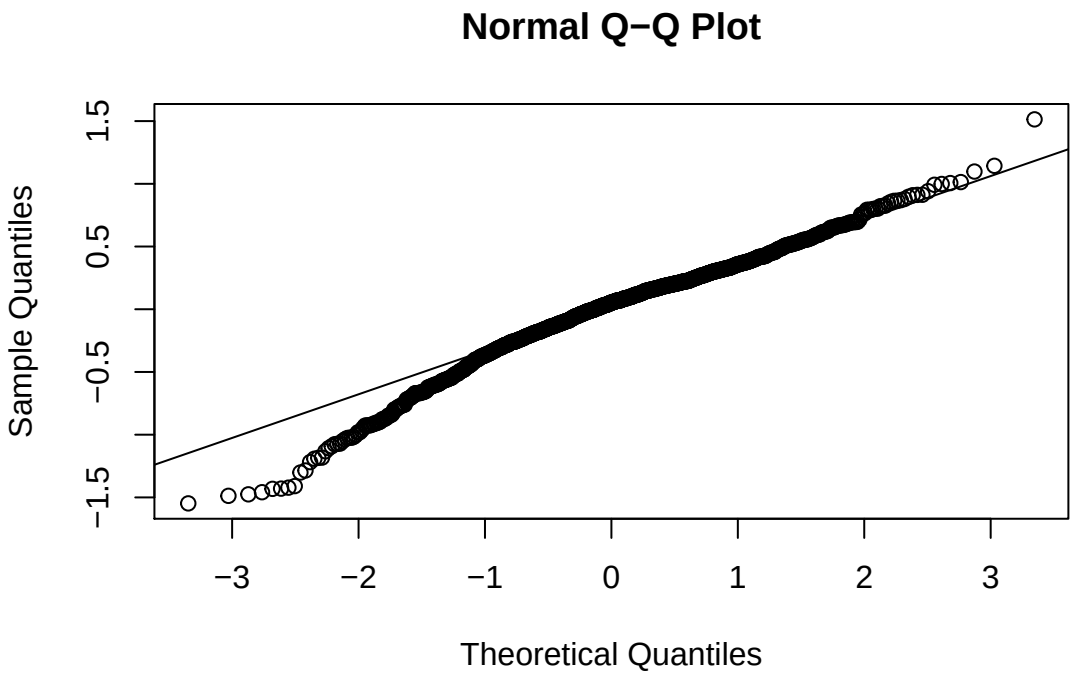
Model Results

For the purpose of learning, I created both normal models and poisson models to investigate their fit. Note that for all of the normal models, the assumption of normality was violated. Additionally for the pooled and stratified models, the assumption of heteroscedacity is violated. Because of this the interpretability is wrong and therefore I would discard all of these models. In the following table, we have a summary of the RSS for the whole dataset and also just for Denver.

Model	RSS_Overall	RSS_Denver
Pooled Linear Model	46788.604	708.3082
Pooled Log−Linear Model	27671.081	324.0425
Stratified Linear Model	28272.684	401.6679
Stratified Log−Linear Model	13804.582	149.0467
Linear Mixed−Effect Model	27433.129	383.2186
Log−Linear Mixed−Effect Model	15148.105	134.0731
Stratified Poisson Model	9403.346	127.4699
Mixed−Effect Poisson Model	9734.990	128.1238
Zero−Inflated ME Poisson Model	149609.976	6245.1368

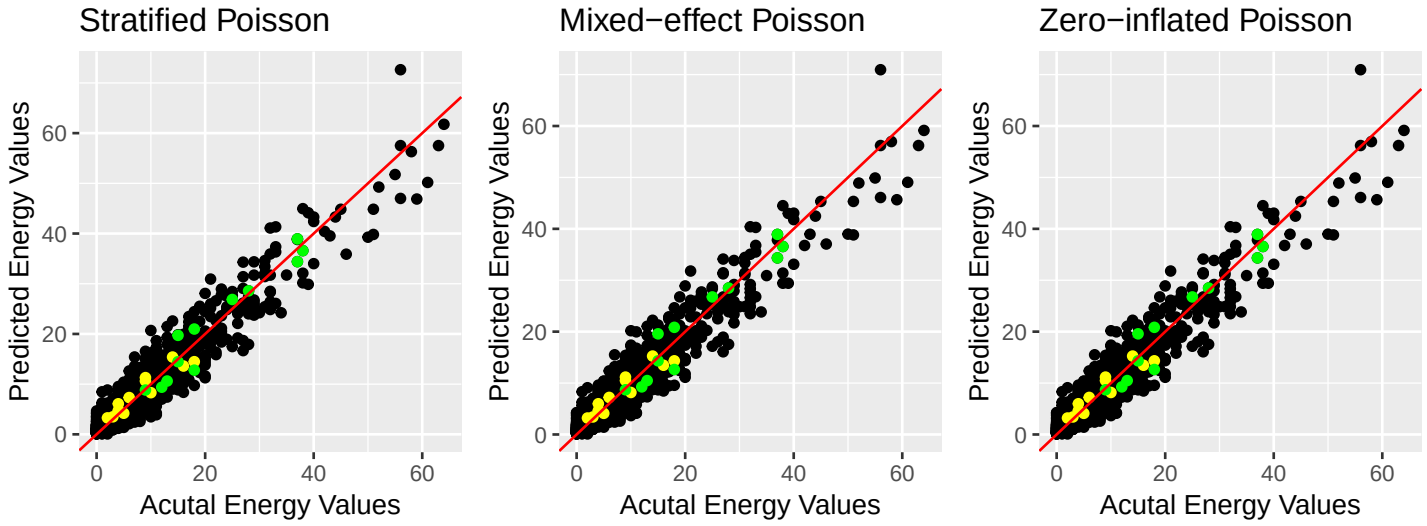
We see that the model that minimizes the overall RSS is the stratified poisson model, closely followed by the mixed-effect poisson. We have the same behavior for the Denver specific RSS. I was surprised that the mixed-effect poisson model did not perform better.

I will note here that many of the assumptions for the normal models were violated. The log-linear mixed effect model did a good job at predicting the energy usage for Denver in the model. However, the qq-plot the residuals clearly indicate that the normality assumption is violated. I decided remove this model from my final consideration. Please see the plot below for justification of this decision.



Poisson Models

There were three main poisson models that I considered using: fixed-effect poisson, mixed-effect poisson, and zero-inflated mixed effect poisson. Here we plot the predicted model values vs the actual model values. The data for Denver is highlighted in yellow and green.



All three models appear to predict the energy consumption for denver well. Therefore I will check the

model assumptions for all three. One of the main assumptions of the poisson model is that the variance and the mean of the distribution are equal. Therefore in order for our model to be valid, this assumption must be met. I calculate the disperson for all three poisson models and include them in the following table.

Model	Dispersion	P_value
Stratified Poisson Model	0.933576	0.0402000
Mixed−Effect Poisson Model	1.030927	0.2206011
Zero−Inflated ME Poisson Model	7.972964	0.0000000

We see from the table above that the mixed effect poisson model has the best dispersion. The stratified poisson model is underdispersed while the zero-inflated poisson model is overdispersed. Therfore, we can conclued that the mixed effect poisson is the best fitting model. However, the p-value for the stratified poisson model is barely significant at the .05 level. I will also consider the stratified poisson model for my final model because although the model is underdispersed, the fit is still good.

Conclusion

For my final model, I chose the fixed-effect poisson because this was the model that minimized the sum of the squared residuals for Denver. I obtained the following predictions:

Month	Prediction_Gwh	LB_Gwh	UB_Gwh
1	16	9	24
2	12.9	6	20
3	9.6	3	16
4	8	3	14
5	6.9	2	13
6	4.4	1	9
7	3	0	7
8	2.8	0	7
9	4	1	9
10	5.6	1	11
11	8.7	4	15
12	12.3	6	20

Note that to compute the prediction intervals for my model, I had to bootstap the training data. My initial concerns about the temperature differences for the two years of data were remedied. The stratified poisson model did a good job at predicting the data values for both yeas of data.

Reflection

From the results of the competition, clearly my results were not optimal. Had I chosen my final model as the mixed-effect poisson model, I may have won the cookies. Perhaps the stratified poisson model overfit the data and caused a biased prediction. My error was in not considering the shared variance for within cities and between cities measurement. This is exactly the situation that the mixed-effect model tries to remedy. Rather than relying on the RSS to choose my model, I should have relied more on measures of model validity. I will certainly consider this in the future.

I tried adding new predictors to the model. I created a categrorical variable for "Region" of the United States. This did not turn out to be effective. I believe that this is due to the fact that Region is too broad of a predictor. I found in the exploratory analysis that all 51 cities have very different energy usage patterns. Surely a predictor of Region (4 regions) will be insufficient for describing the trend of the data. Therefore I excluded the reegion predictor from my model. I also created a continuous variable that represents the population of the city. I included this predictor in all of my normal-models; the predictor was insignificant in all normal models. Additionally I tried including this predictor in my poisson models. It tried using population as an offset term for my regression. This too turned out to be ineffective. The reason for this is that a cities population changes dynamically. It is possible that our two years of data were taken from very distant years, allowing for a population change in the United States. Perhaps if the two years are close to each other, the offset of population should be considered in the poisson models.