Practicum Fall Summary of Work

Project Overview

Executive Summary:

In this project, economic, demographic, landuse and weather data were used to predict crime rates at the census tract level. This data was collected from a wide variety of sources and processed such that the monthly crime rate for a given census tract could be predicted. Initially, basic models were tested (linear regression, Poisson regression, decision trees), and although the models had good predictive accuracy they had no provision to account for the variance of crime at the individual tract level. This gave motivation to using bayesian models to account for the high variance in crime rates. First, a bayesian linear regression was run and plotted which showed to encompass a majority of the variance in the test data. Next, the economic and demographic data was used to cluster the census tracts into four groups.

For next semester, I plan to use these generated groups as well as other geographically created groups to implement a hierarchical model structure. I also plan to explore using bayesian analysis for the Poisson model as well as using BART for the tree models.

Proposal:

Classical machine learning models have been applied to predicting crime. However, crime rates generally have a high variance, which is motivation to use a Bayesian analysis to predict and model crime. I am currently using demographic, economic and business data in the analysis, as well as monthly trends such as temperature and precipitation. With this data I am applying standard models (Linear regression, Poisson regression, Decision Tree) and applying different Bayesian structures to each of these base models.

Problem Definition:

The main analytics question is for a given census tract predicting the crime rate by month. In other words, how accurately can we predict the number of crimes per month in a given census tract.

Data Processing

Data Collection:

The main data collected in this project was Philadelphia demographic, land-use and economic data. Philadelphia crime data and land-use data was used from the paper, “ANALYSIS OF URBAN VIBRANCY AND SAFETY IN PHILADELPHIA”.

The demographic data was collected from the Decennial Census API. This api-script had to be re-created from the above paper since the API has changed. For all code please refer to the github:https://github.com/adisrivatsan/Philadelphia-Crime-Analysis-Practicum . Similarly, the economic data was collected from the ACS Census API. The data was collected at the census tract level as this was the main granularity of analysis.

In terms of the vibrancy/business data, an attempt was made to gather the data from Yelp, Foursquare and Google Places, however, not enough data was collected. Now the data is in the process of being collected.

The weather data collection process was interesting since there was no straightforward API to get historical Philadelphia weather data. A web scraper was created to get the relevant data for the relevant time period from https://www.wunderground.com/history/.

For more detail in this process please refer to IPytonNotebooks/Practicum Data Load.ipynb, and IPytonNotebooks/Weather Data Import.ipynb in the repository.

Data Cleaning:

The data cleaning process first involved defining a good data structure based on the problem statement. Since the problem statement above looked at number of crimes per month in a given tract, the primary key was (tract,month,year). This process involved merging many tables and filtering the data such that the data was aggregated at the correct level. Furthermore, many of the variables were categorical and was therefore, converted into numeric variables appropriately. This process was a fairly time consuming process.

For more detail in this process please refer to the: IPytonNotebooks/Practicum DataCleaningPythonProject.ipynb in the repository

Exploratory Data Visualization and Analysis:

During this phase of the analysis, different aspects of the data were plotted to get an intuition on possible predictive algorithms. First, the number of crimes by type was plotted and can be seen in Figure 1 in the appendix. Next, the number of crimes by year was plotted from 2006 to 2015 seen in Figure 2. There is a decreasing trend in crime rates which confirms initially using a linear and Poisson model. Then the number of crimes by month were plotted in order to identify different types of seasonal trends seen in Figure 3. There exist seasonal trends and therefore, temperature and precipitation data were included to account for these trends. The next three figures include the number of crimes per year for each crime type seen in Figures 4 through 6. On a high level, all of the crime types display similar trends and therefore, all crime types were modelled similarly. An interesting direction to take in the future is to apply these models to predicting individual crime types.

For more detail in this process please refer to the: IPytonNotebooks/Data Preprocessing and Visualization in the repository.

Data Analysis

Preliminary Analysis:

In order to identify the predictive power of demographic and economic data on crime, a small analysis was performed in which the dependent variable was the number of crimes in a tract and the independent variables were the demographic and economic data. This analysis was performed to simply test the significance of the demographic and economic variables in predicting the level of aggregate crime in Philadelphia census tracts. Figures 7 and 8 show a linear and random forest fit of the data.

The overall result is that economic and demographic data are significant variables to predicting crime. As this analysis is not the main analysis of the project, more description about the methods and results can be found in the related notebook.

For more detail in this process please refer to the: IPytonNotebooks/Clean Data Analysis in the repository.

Basic Model Analysis:

Now since the data has been collected, cleaned and processed, basic analytical models can be run and evaluated. First the data was split into the training and testing by year. Years 2006-2013 were in the training set while years 2014-2015 were in the testing set.

The main three models tried were: linear regression, Poisson regression and Decision Tree. Figures 9 and 10 below shows a fit of these three models on a sample census tract for both the training (in-sample) and testing (out-sample) sets. Overall the linear and poisson models performed similarly and the decision tree performed worse.

Next, regularized linear models were tested such as Lasso and Ridge regression. Figures 11 and 12 below show a similar fit of these models to the in-sample and out-sample data.

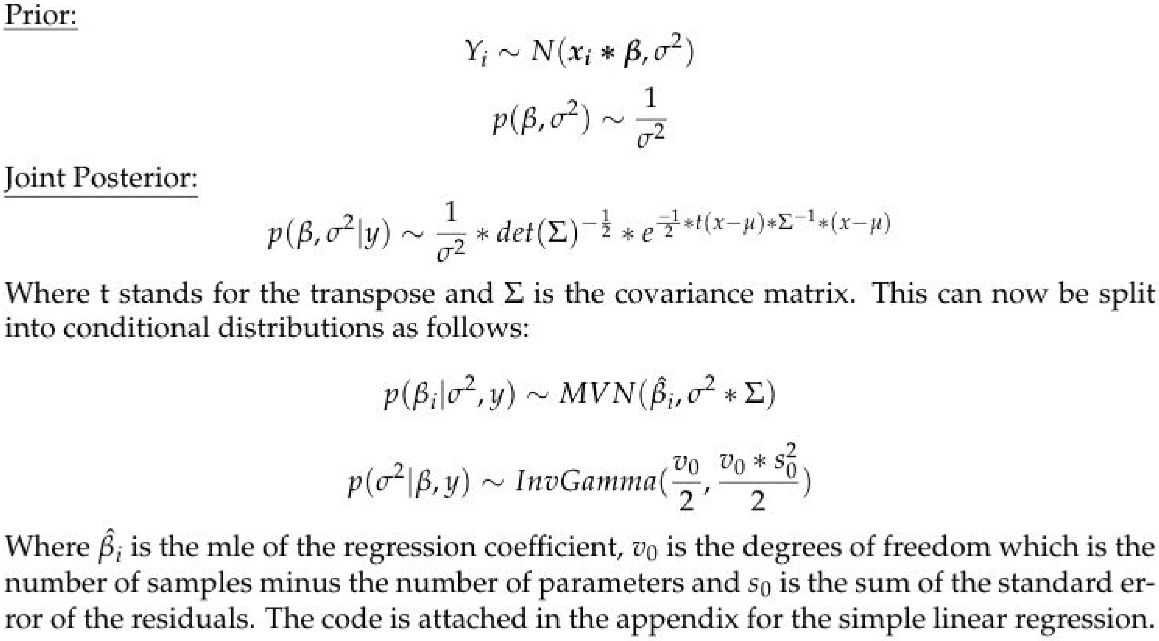
Basic Model Accuracy Testing

The accuracy of the above models for all the tracts were tested. Figures 13 and 14 below show the mean squared error accuracy of each of the five models described above in both the in-sample and out-sample data for all tracts. The results were that the Decision Tree model performed significantly worse than the other four models and all of the linear models performed roughly the same. To reiterate a separate model was fit for all of the 384 census tracts and the average mean squared error is displayed below.

Although these models generally have a good fit, they could be overfitting the data as well as not accounting for the in-built variance of crime rates at the individual tract level. Figure 15 displays the out-sample model variance for all five models. These models only have a maximum variance of around 50 while the actual variance is 150. This is the motivation to explore bayesian models.

Bayesian Linear Regression

Since the linear and Poisson models had similar accuracy, it would be interesting to see if they also have similar results when a basic bayesian prior is applied to the parameters. The bayesian linear regression model is described below. This is a basic version of the model and other structures will be explored next semester.



Prior:

$$Y_i \sim N(\boldsymbol{x_i} * \boldsymbol{\beta}, \sigma^2)$$

$$p(\beta, \sigma^2) \sim \frac{1}{\sigma^2}$$

Joint Posterior:

$$p(\beta, \sigma^2|y) \sim \frac{1}{\sigma^2} * det(\Sigma)^{-\frac{1}{2}} * e^{\frac{-1}{2} * t(x-\mu) * \Sigma^{-1} * (x-\mu)}$$

Where t stands for the transpose and Σ is the covariance matrix. This can now be split into conditional distributions as follows:

$$p(\beta_i|\sigma^2, y) \sim MVN(\hat{\beta}_i, \sigma^2 * \Sigma)$$

$$p(\sigma^2|\beta, y) \sim InvGamma(\frac{v_0}{2}, \frac{v_0 * s_0^2}{2})$$

Where $\hat{\beta}_i$ is the mle of the regression coefficient, v_0 is the degrees of freedom which is the number of samples minus the number of parameters and s_0 is the sum of the standard error of the residuals. The code is attached in the appendix for the simple linear regression.

In the above description, Y_i refers to the number of crimes per month for an individual tract. The x_i and beta_i variables are the covariates.

Using this process, 1000 regression coefficients were sampled from the above distribution. These coefficients were then used to generate 1000 different linear regression predictions for each tract. Figures 16 and 17 display all 1000 generated lines on the sample tract used above. Notice how the variance of the model plays a role in determining the shape of each line. Notice, also how the variance changes between the in-sample and out-sample data.

Bayesian Model Evaluation:

A way to evaluate bayesian models is to use the posterior predictive distribution to generate data. The 1000 regression coefficients were then used to generate data for the out-sample set for all tracts. For each set of generated data the correlation between the generated data and the actual data was used to evaluate how closely correlated each generated data was to each tract. Correlation was used as a metric in this scenario since it is important to understand when crime rates increase and decrease. Variance was also measured as a metric for the model since the main motivation for using bayesian modelling was to be able to capture the variance in crime rates. Figures 18 and 19 below, show the distribution of the average correlation and variance between all of the census tracts.

Interesting results show that the correlation is roughly around 0.5 which is good. Also the variance distribution shows a large concentration below 100 but a small mass of concentration around 100-200. The mean variance of the actual test data is 150, which means that this model is capturing the average variance of the test data. As these models are iterated on, hopefully more of this variance can be captured in this posterior predictive check.

K-Means Clustering

Since the economic and demographic data does not change over time and this analysis is performed on the month interval, the economic and demographic data was used to create different clusters using K-Means. Solely using the demographic and economic data, each of the tracts were split into 4 clusters, which was empirically decided by graphing the distortion of Kmeans shown in Figure 20. Figure 21 shows the number of tracts per cluster. This type of clustering is the first step in building models that share information between clusters.

For more information on the above process please refer to: Data Analysis Main.R in the repository.

Next Semester Work

To recap, the data has been cleaned, processed and analyzed at a basic level. Next semester, I hope to use these clusters generated from KMenas to build more complex hierarchical models. Similarly, I want to explore hierarchical Poisson and BART models as well to identify if the bayesian versions of these models can add predictive power.

As stated above, the main goal of this project is to capture the variance of crime rates in Philadelphia, and use these models to simulate certain crime scenarios in order to better prepare law enforcement to handle these situations. As more models are tested, hopefully the predictions will become more accurate and the simulated scenarios more realistic.

Appendix:

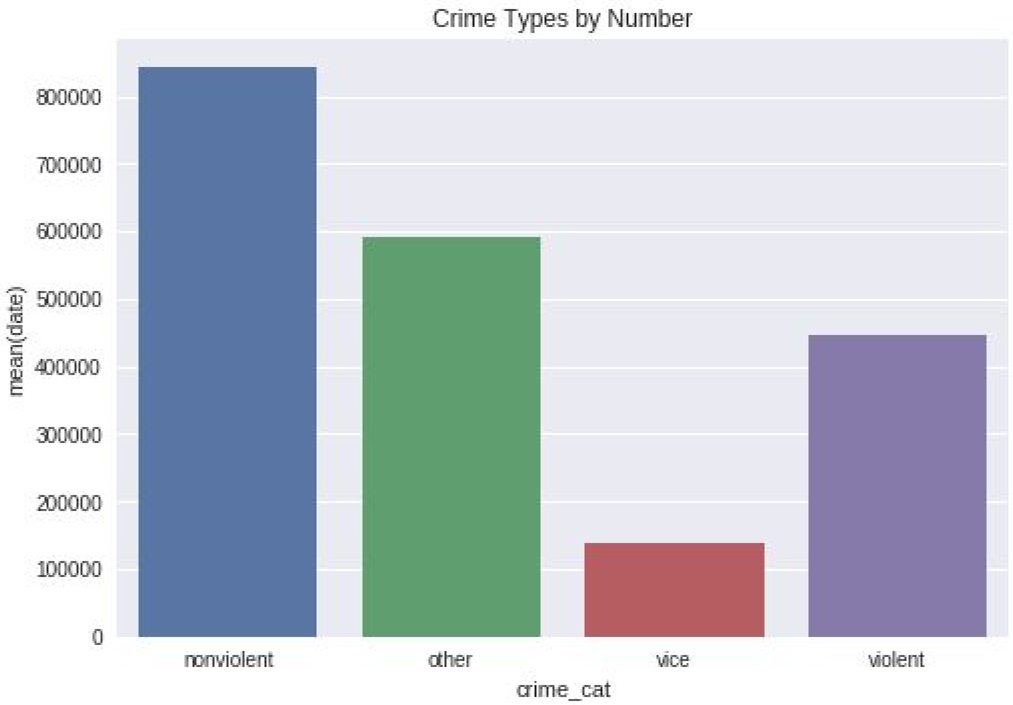


Figure 1: Crime Type by Number

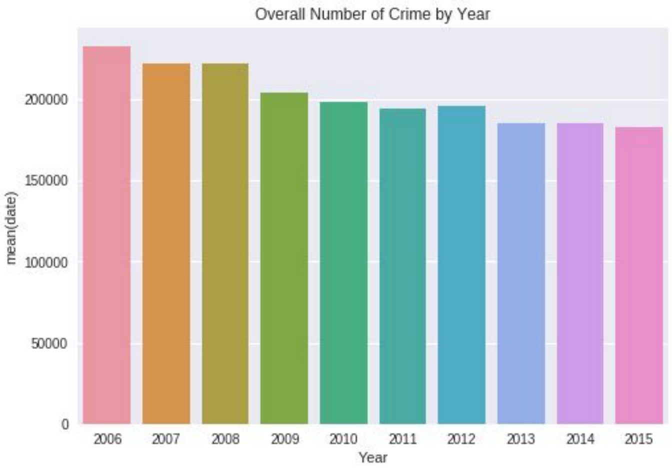


Figure 2: Number of Crimes per Year

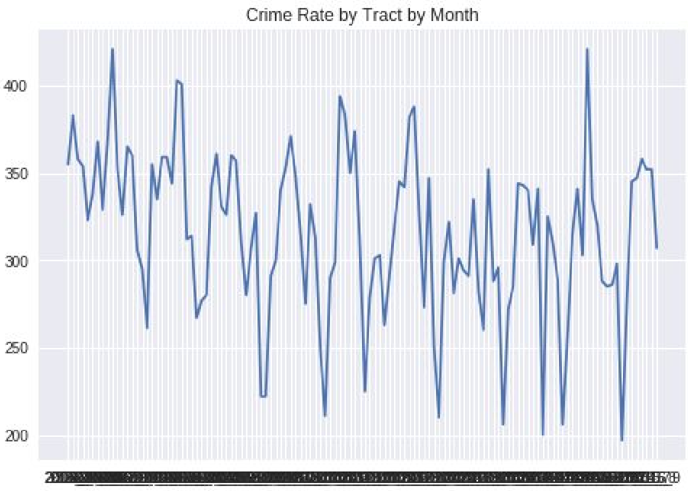


Figure 3: Number of Crimes per Month

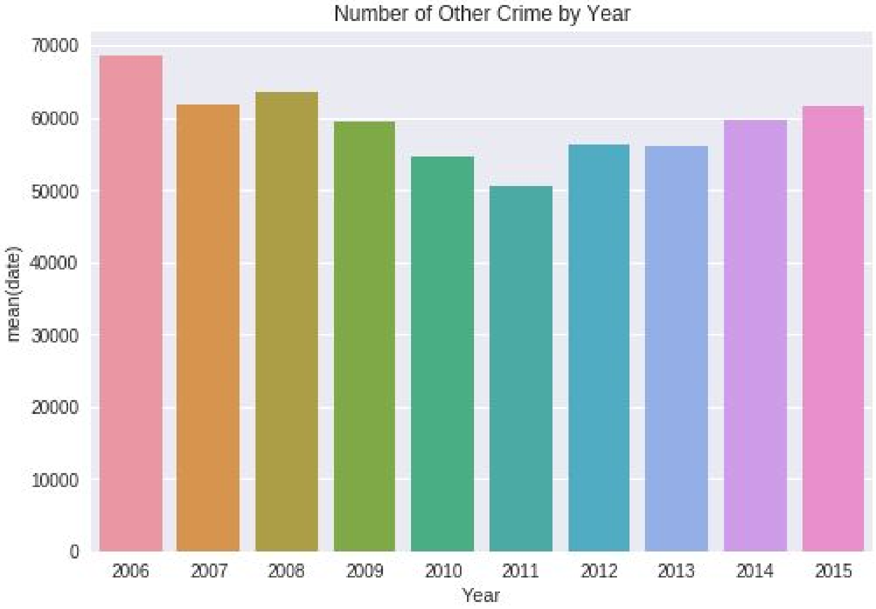


Figure 4: Number of Other Crimes per Year

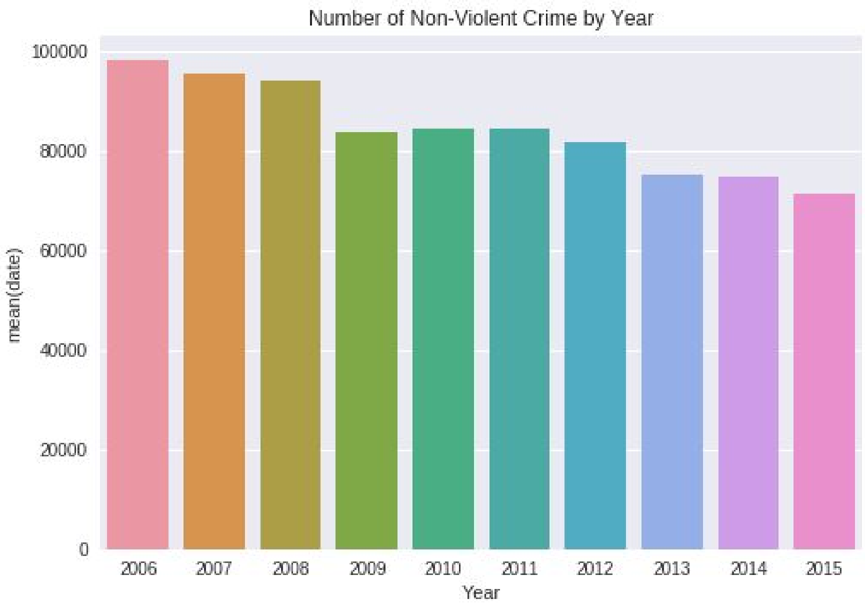


Figure 5: Number of Non-violent Crimes per Year

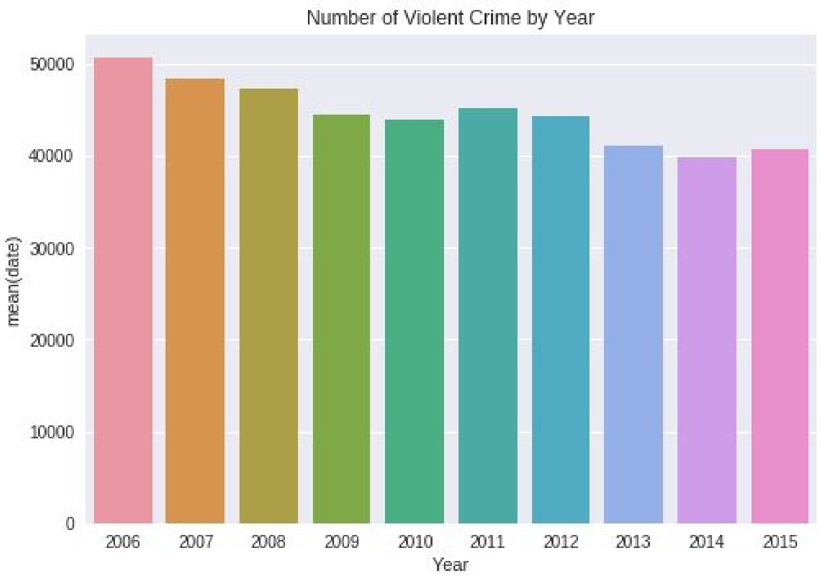


Figure 6: Number of Violent Crimes per Year

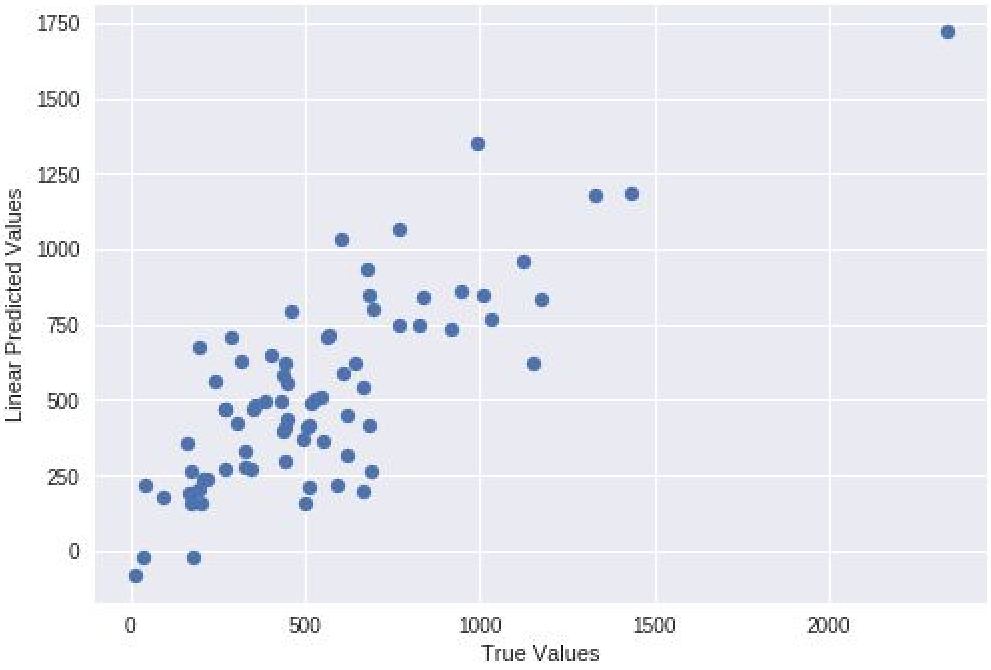


Figure 7: Linear Model Fit

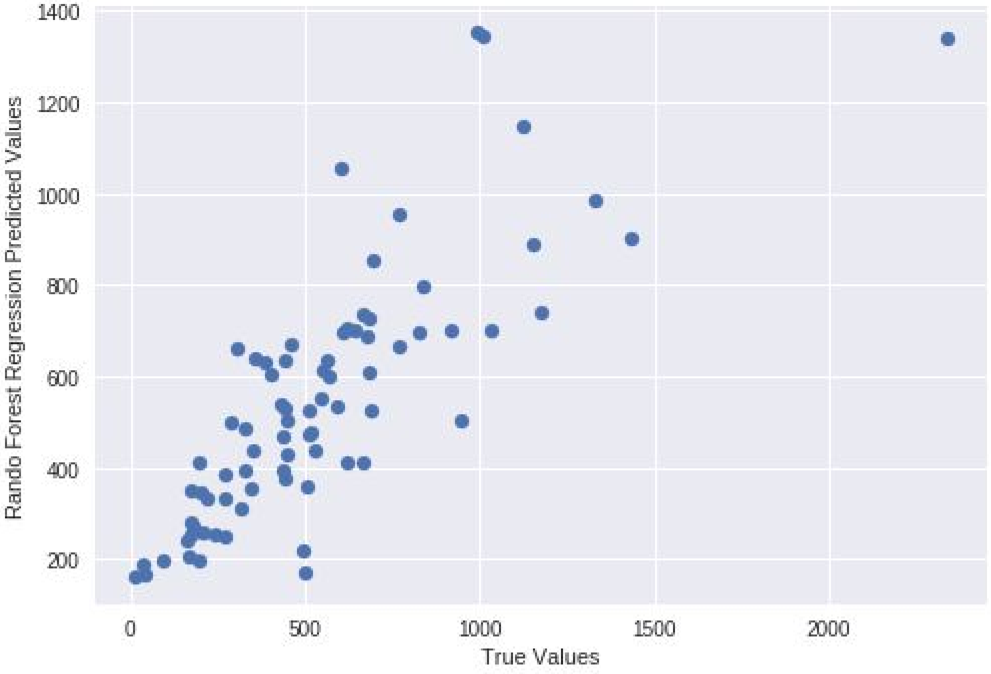


Figure 8: Random Forest Model Fit

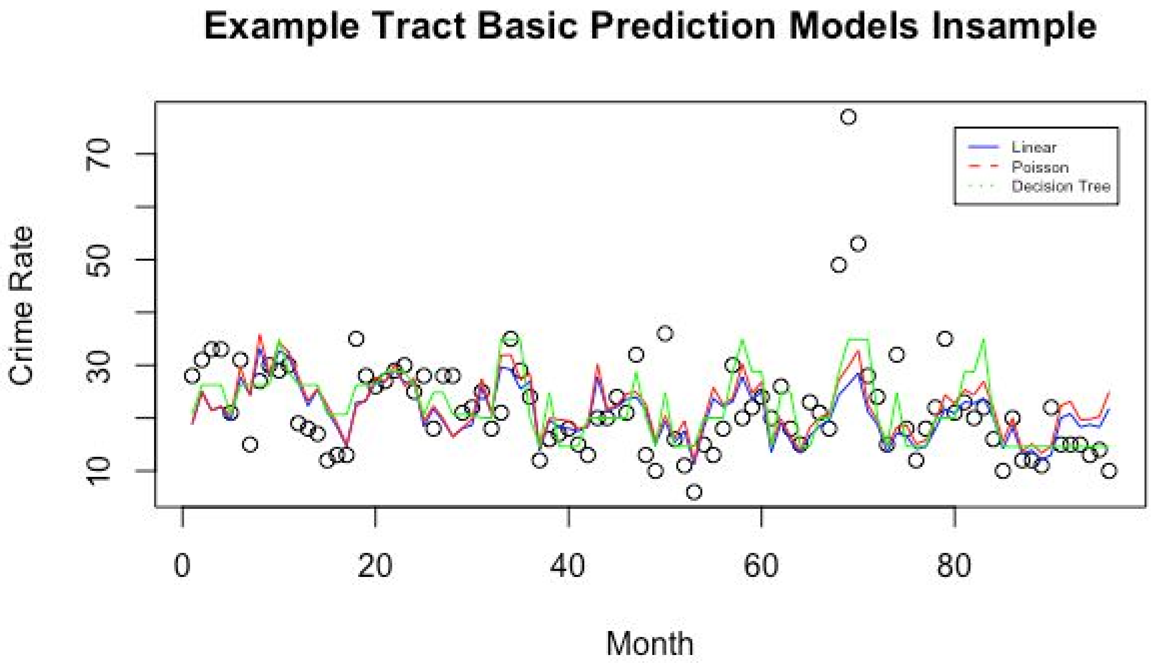


Figure 9: Basic Predictions on Training Set

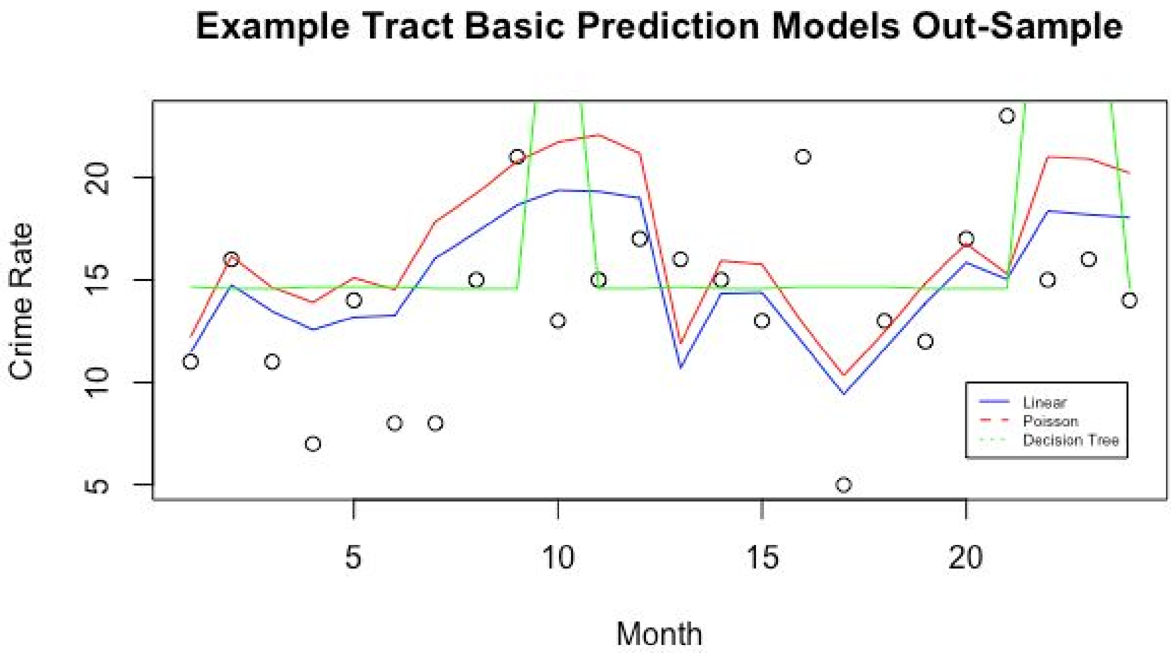


Figure 10: Basic Predictions on testing data

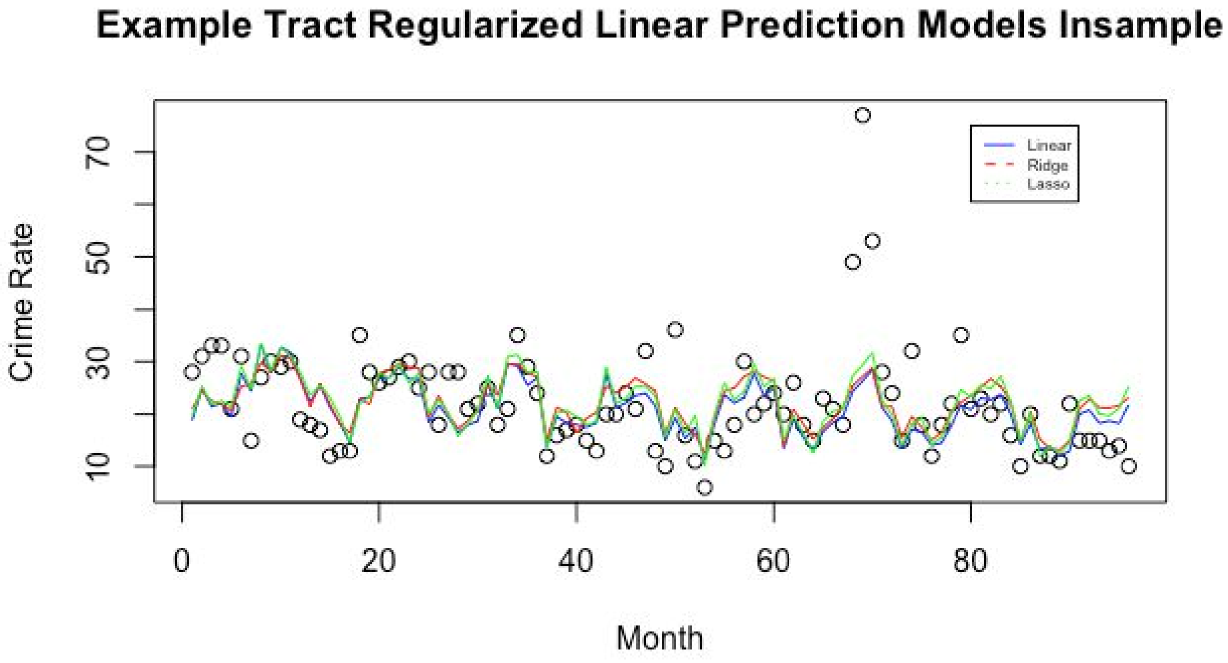


Figure 11: Regularized Predictions on training data

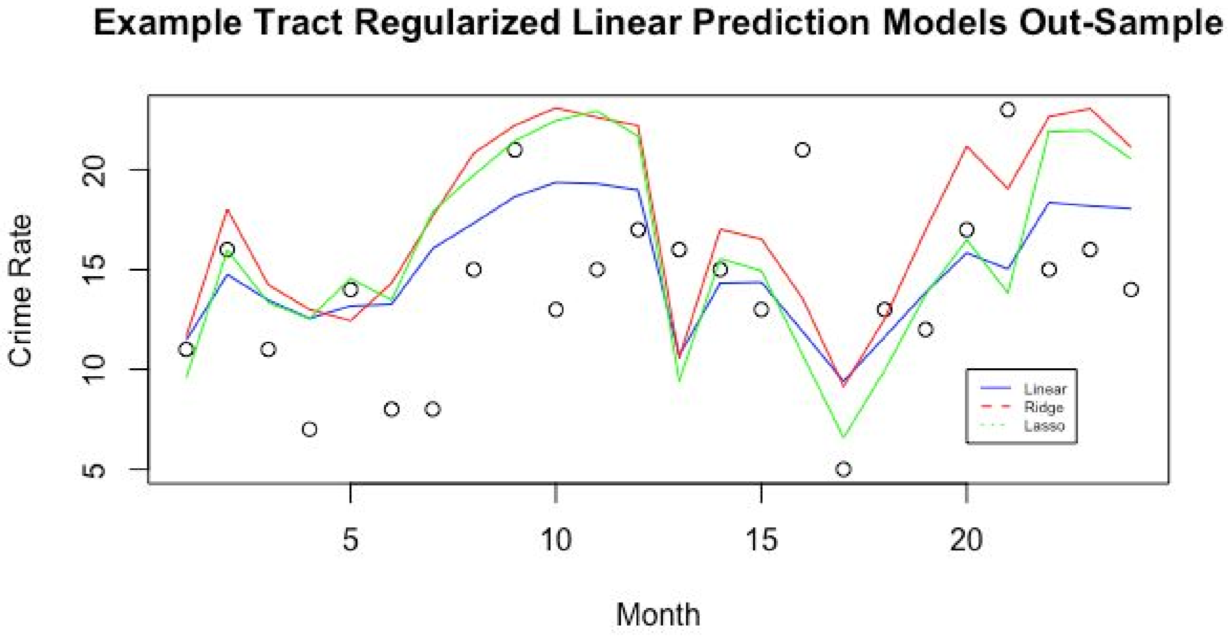


Figure 12: Regularized Linear Model Out of Sample

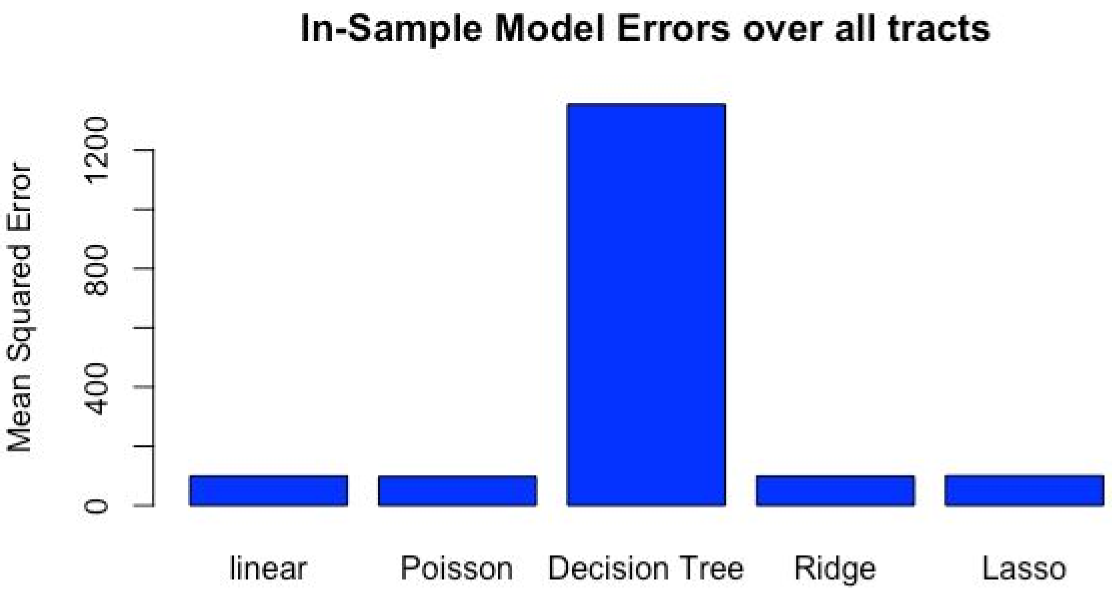


Figure 13: In sample model errors

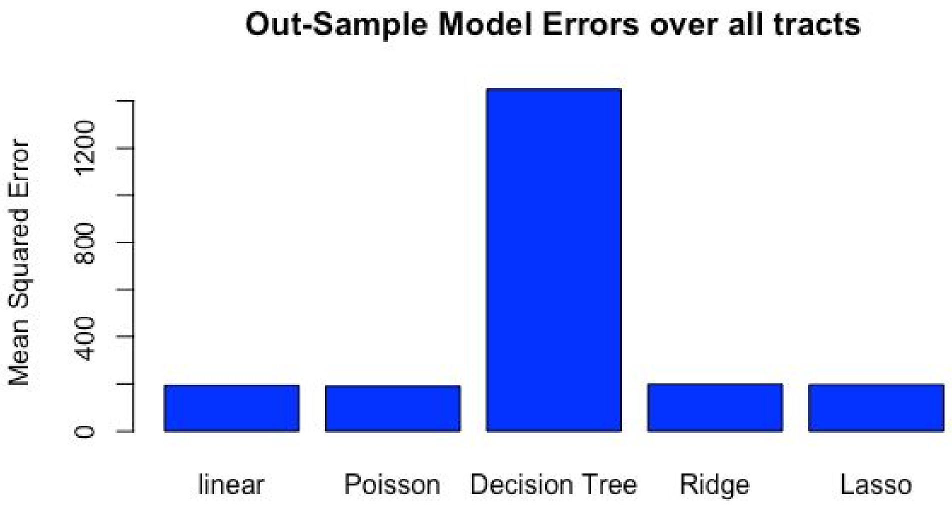


Figure 14: Out-sample model errors

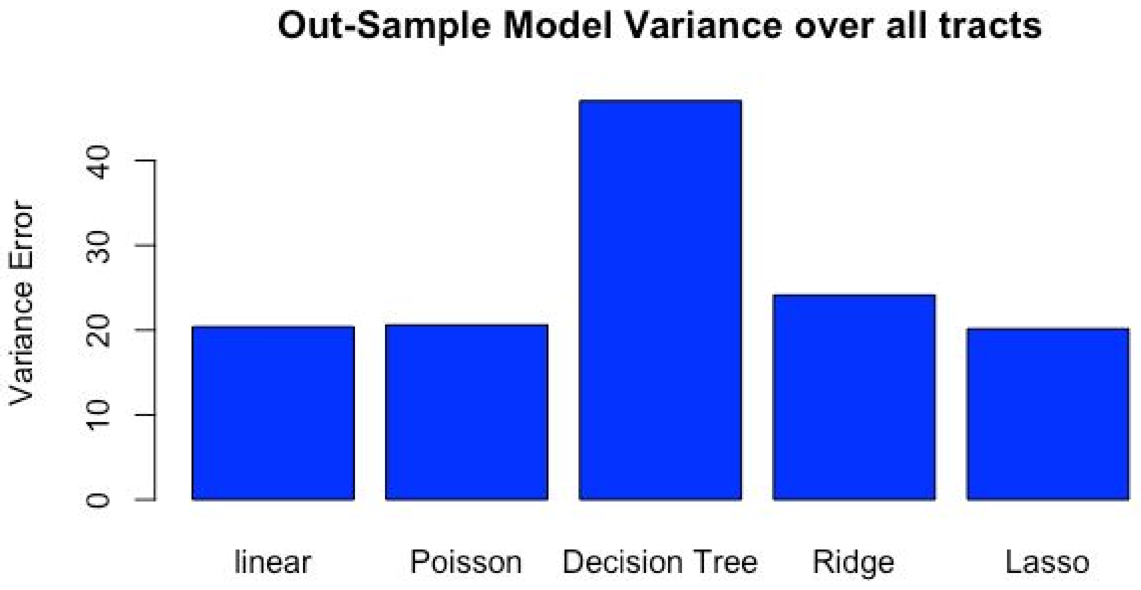


Figure 15:Out-Sample Variance for all Models

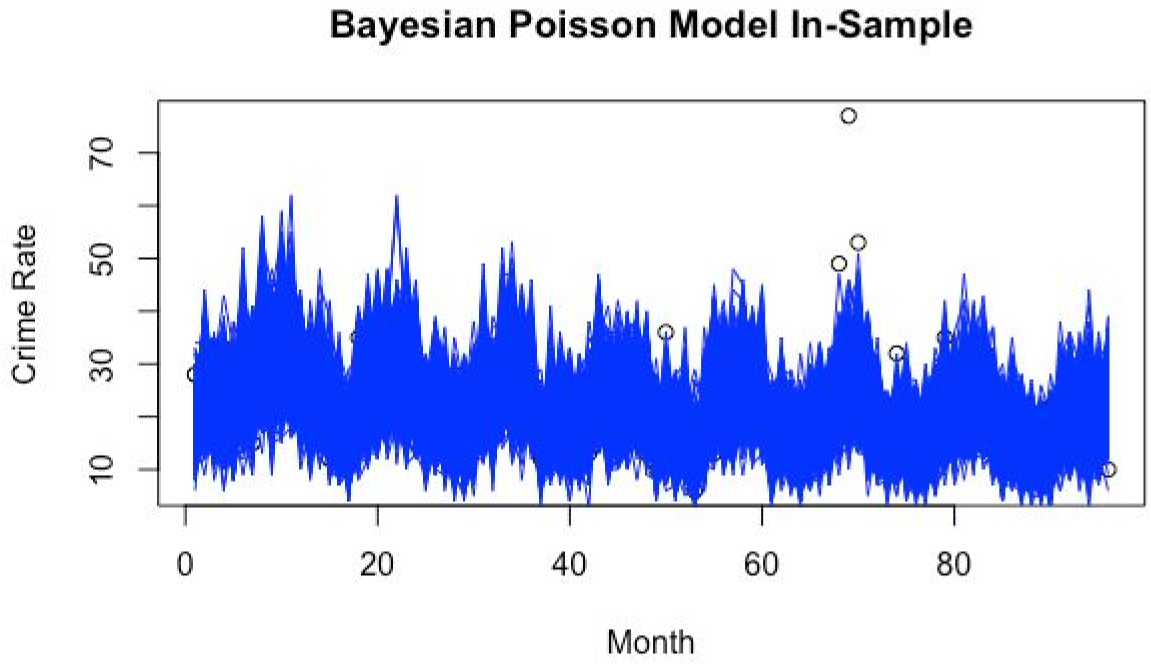


Figure 16: Single tract In-sample Bayesian Model

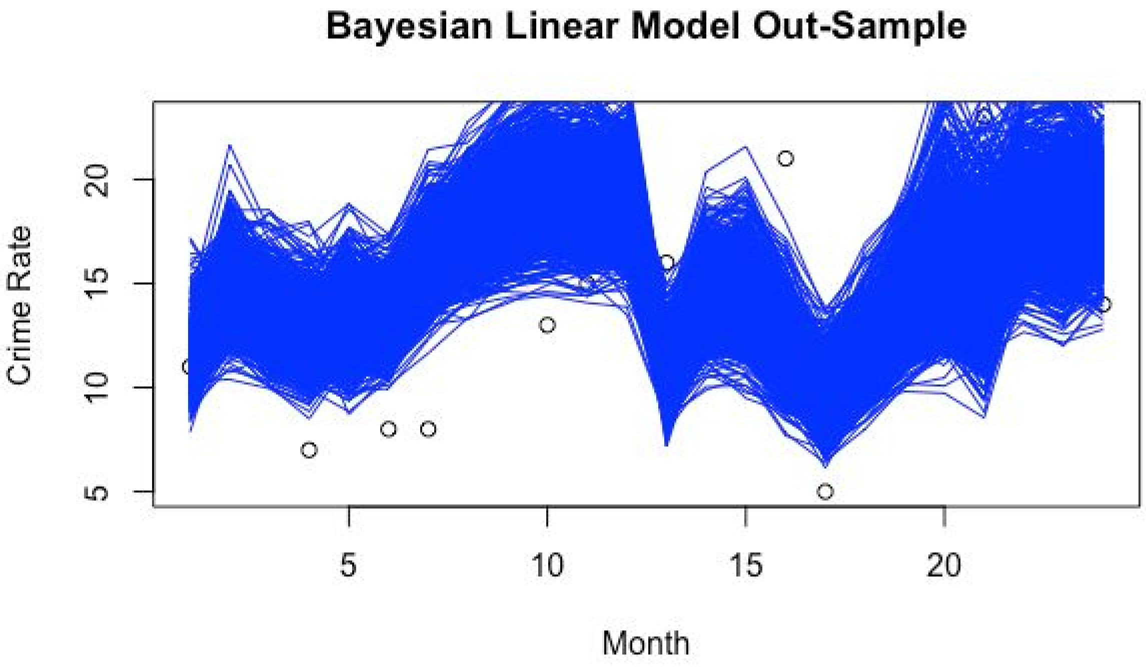


Figure 17: Single tract Out-sample Bayesian Model

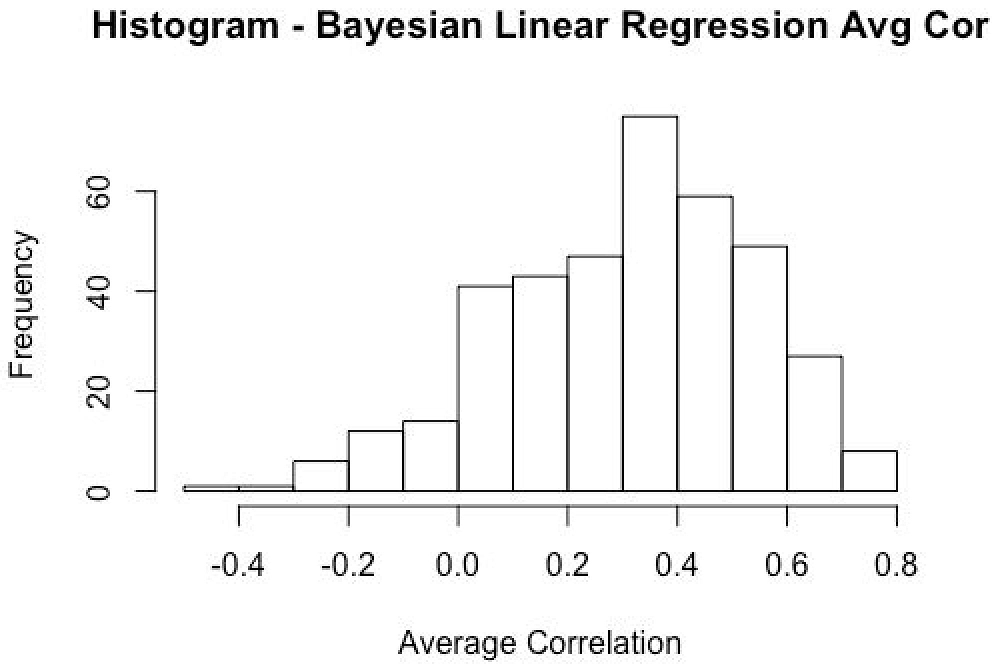


Figure 18: Average Correlation over all tracts

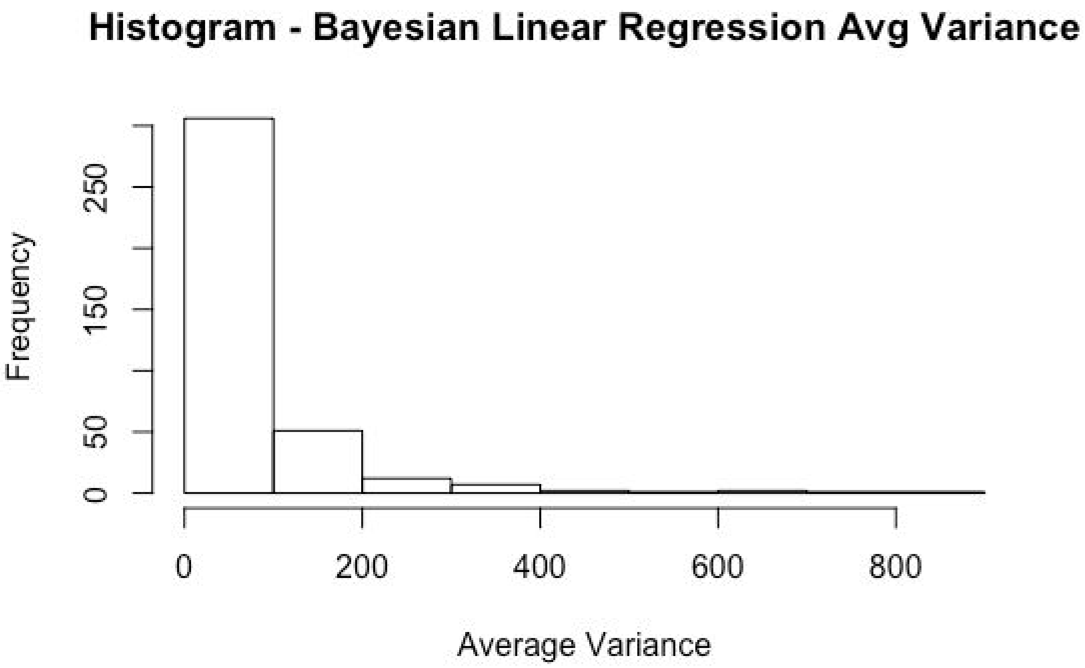


Figure 19: Average Variance over all tracts

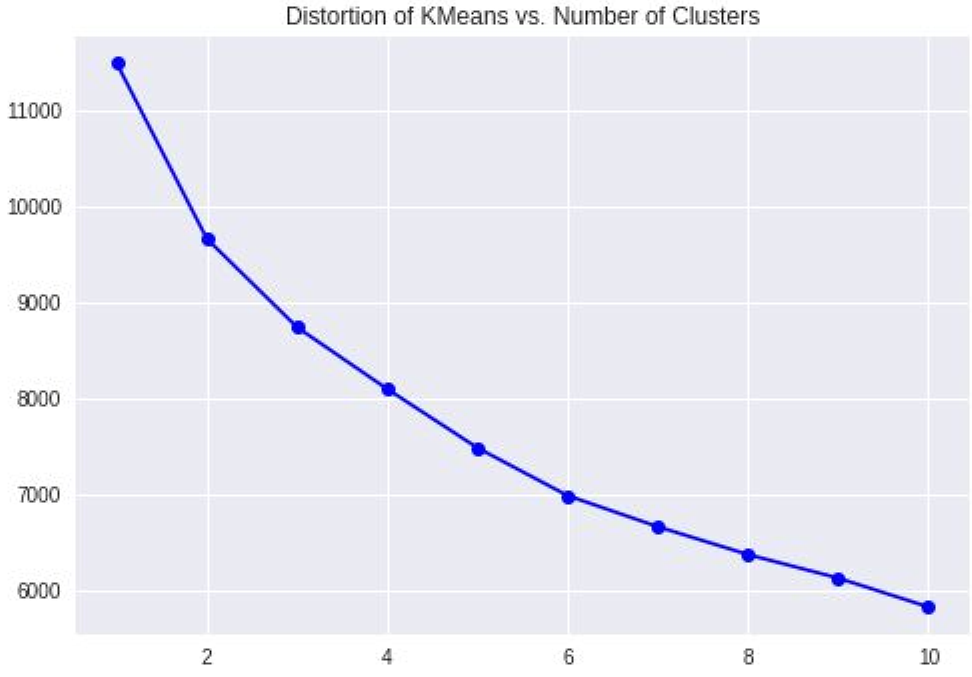


Figure 20: Distortion error in trying different number of clusters (elbow is around 3-4)

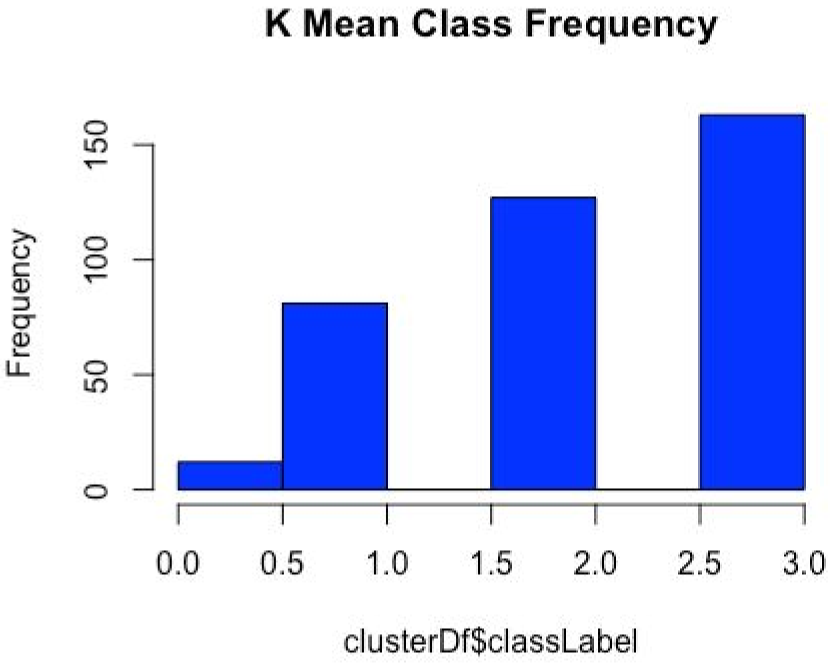


Figure 21: K Means Class Frequency