# **Public Policy Question Assignment**

Kevin Coleman

# Introduction

For this assignment, you will write a proposal synthesizing the information that you have learned to design an analysis to support data-driven decision making in public policy. In this assignment, you will not be conducting the analysis, but you will be designing your "ideal" plan. Propose a question that is relevant for local, state, or federal government, identify the information needed (data source or data sources) to conduct an analysis, and propose the analytical strategy to be used. Choose the model covered during the semester that best fits your analysis (a decision tree, random forest, hierarchical cluster analysis, or k-means cluster analysis). You must choose a data source *other* than the data that we have used during the semester.

## Directions

Any of the lecture and reading material for the semester may be useful for the assignment, but several readings in particular will be good resources. This includes *Practical Data Science with R*, Chapter 1, "Chapter 3: What is the Question?" *Research Methods for Public Administrators*, "Brilliant Analytics for Smart Cities," *A Practical Guide to Analytics for Governments: Using Big Data for Good*, and *Practical Data Science with R*, Chapter 5, Section 5.1, Chapter 11.

The completed assignment should adhere to the following guidelines:

* 1. Include your answers on the assignment document. The answers should be in essay form.
  2. Write your answers using complete sentences with correct punctuation, grammar, and spelling.
  3. Submit your completed assignment through the Blackboard portal in Lesson 15

Answer the following questions:

1. What is the public policy question? (2 points)

The public policy question will focus on housing and monthly rent costs within Pittsburgh, PA. Specifically, the public policy question will look into the greater Pittsburgh metropolitan area, represented by Allegheny County. Elected officials would be concerned about their constituents housing/rent costs. The county of Allegheny’s Housing office may ask: What factors affect housing/rent prices in certain towns and neighborhoods? How do average housing/rent costs vary in different towns/ neighborhoods? The Allegheny Housing office wants to know which socio-economic variables are associated with housing/rent costs. The public policy question is a descriptive type question that seeks to understand the housing situation. The goal is to provide the basic facts before any kind of additional analysis takes place. It is possible that future questions may lead to more relationship style questions such as, what is causing rent/ housing to increase?

1. What are the data sources that you would use to conduct your analysis? Identify the source of the data (the organization that produced the data and where the data are found). Describe the characteristics of the data including the unit of analysis, sample, and variables. (2 points)

The data source I would use would come from the US Census Bureau, specifically the ACS 2020 dataset. The American Community Survey is a credible, reliable, and free public database that contains the variables of interest for the policy question. The year 2020 would be used, since the most complete information is collected at the Census. By using the tidycensus package in R, the required data for Allegheny County is easily accessible within the R environment. The unit of analysis will be at the tract level, since ACS data is aggregated this way and it matches well with observing neighborhood/ community trends. ACS data is already a sample of the population as it surveys a subset of households for its data collection. The dependent variable would be a combined housing cost variable composed of rent cost and house owning costs. The predictor variables would consist of socio-economic variables such as median income, education level, ethnicity, marital status, and number of children. The predictors could be further split into additional sub variables such as white, black, and Hispanic for the ethnicity variable. Education level can be split by no high school, high school, and college educated. Marital status can be split by single, married, divorced, and widowed. The predictor variables are counted at the household and individual level, but the analysis would occur at the tract level to avoid ecological inferencing between the predictors and the dependent variable.

1. Which machine learning model would you use and why? Describe the stages of your analysis from the original assessment of the data through the analysis, and the subsequent model diagnostics. (4 points)

I would use the random forest model. A Random Forest is a more complex analysis of decision trees and uses repeated random samples of data and variables from many decision trees that creates a well generalized model. A random forest is essentially many different kinds of decision trees. Due to this, the random forest model benefits from not needing to prune or change the parameters of an individual tree. Since this model will contain count, continuous, and categorical variables, random forest is able to handle the different types of variables combined into one model.

Before the model is created, the data would need to be processed. This includes hot encoding categorical variables such as marital status and race/ethnicity. Depending on the amount of potential missing data, I would likely use a mean or median imputation for continuous variables like income, however it may have to be localized since a neighborhood with a lower income would likely not yield a median income for the entire city. However, the random forest model should be able to handle some random missing data. I can also create interaction terms to see the combined effect of multiple predictors on the target variable. For example, I would likely make an interaction term of income and education level for housing/rent costs. I would also make another one for race/ethnicity and income on housing/rent costs. These interactions appear to be aligned, and can be tested by running these terms into the model. Random forest conducts feature selection through its bagging and variable importance processes. I would not create manual subsets. Scaling should not be a concern as well, since random forest models make splits in the data based on determined values within the various decision trees.

The random forest would be created by using the rattle GUI since it provides an easy interface that can easily run a random forest model. To begin, the random forest model conducts ‘bagging’ where it samples usually about 70% of the total observations. The other 30% of the data is called out-of-bag estimate or OOB. OOB shows the error rate or how the model performed when that 30% of the data was not used in the random sample. Random forest conducts its own train/test split as well as cross-validation. Additionally, random forest shows the error rate, either RMSE for continuous variables or the misclassification score for a categorical variable, and the within-class error rate which provides a more localized insight into where the model is making mistakes such as a certain percentile range. For example, if the random forest model gave a OOB error rate RMSE score of 150, the ‘test’ data would be $150 off of the ‘training’ data used in the model. If the RMSE score, from the ‘training’ data, scored 50 than it would be $50 off from the actual housing cost. Both the RMSE score from the training data and the OOB error rate RMSE are relatively low compared to overall housing costs. This suggests that the model is generalizing both unseen and current data well for its predictions.

Random Forest can also sort predictor variables by variable importance. This can provide insights as to which predictor variables are most important in affecting the target variable (housing cost). A variable with a higher score is more important than a variable with a lower score. Random forest conducts its own internal splitting, cross-validation, error rates, and variable importance score. It is an extremely robust model that can tackle many different kinds of variables while accounting for possible overfitting issues as seen with decision tree models. Random forest can be computationally challenging due to the number of trees, variables, and the size of the dataset. However, assuming that I worked for the federal or local government, I should have the required resources and time needed to create and run a random forest model. If time and resources were a concern, I would instead create many different kinds of random forest models for each tract or set of tracts rather than the whole county of Allegheny.

1. What information would your analysis produce that a policy maker could use to make decisions about the public policy question that you have identified? (2 points)

The main goal of the machine learning model was to ascertain the factors that are influencing housing prices and to observe how average housing costs change. By observing how the predictors interacted with the target variable and predicting where housing costs will increase or decrease, policy makers can identify which socio-economic variables have the most important shared relation with housing cost and could effectively target communities with more expensive housing costs rather than waste resources on communities that have low housing costs. For example, if median income was determined to be an important variable in predicting housing costs, policy makers could enact or change policy affecting median income and housing prices. If neighborhood A was predicted to have very high housing costs compared to neighborhood B, policymakers could entice realtor developers to build more homes in neighborhood A. This would theoretically lower the price of housing in neighborhood A. Or, if a business park was opening and the stake holders wanted city officials to recommend a location, the officials may recommend a site near neighborhood A. This could raise the median income of neighborhood A, who would be better able to afford housing/rent costs. Obviously, there are many variables that play into housing, but it can at least provide some kind of framework for city planners and policy makers to make more informed decisions.

## Scoring

The assignment is worth 10 points. Points will be based on the real-world usefulness of the question, the appropriateness of the analytic strategy for the question, and the thoroughness and attention to detail of the answers.