Michael Perrine

DSC 550 Data Mining

Professor Werner

Final Assignment

Analysis of Population Data

Milestone 1

What do demographics, economic growth and school ratings have in common? These factors are important to developers. The cost of building in the wrong location can cost millions in losses. The goal of this analysis is to determine locations of interest for land development. We believe Texas is an attractive place to develop land for the purpose of economic growth. To validate this assumption, I will explore several factors that make a location attractive. These factors are the quality of schools, economic development, and population growth.

To perform my analysis, I will use several libraries found in the python program. Some of the libraries are Pandas, Matplotlib, NumPy, Scikit learn, and Seaborn among others. These libraries will give me the ability to perform my preliminary analysis. The data will come from the Texas Open Data Portal: State of Texas | Open Data Portal | Open Data Portal. This website has csv files for all the data of interest. To begin the project, I will upload the files into my Python program. I will need to perform some data analysis to uncover relationships such as what factors stimulate population growth? What factors drive business? What impact does a good school play in migration decisions? How can I use this information to predict areas of growth?

Some of the work that I will perform will be exploratory data analysis. I will need to transform and clean the data sets. Once I clean the data it will be ready for the visualization process. Visualizing the data is another important tool to analyze and interpret data. This includes building graphs such as histograms, scatter plots, and box plots. These are just some of the visualization tools available.

After I build my graphs, I can determine the relationships involved. The next step is to build a model. Scikit Learn is the library that will build the model. This is probably the most important step in the process because it gives me the ability to make predictions and take actionable steps in determining what areas are prime investment opportunities.

There are challenges to my analysis. I don't know what relationships if any I will find. If I do find relationships in the data, I'm not sure how strong they are or if they will lead to a viable investment decision.

In [ ]: