**Data Science Project**

****

**Data**

The dataset for the project consists of U.S. Census data obtained for over 2K U.S. cities. It also contains 2012 state election results and annual average high and low temperatures for each state.

**Goal**

The goal of the project is to perform data mining on the U.S. census data in order to discover interesting data patterns. For example, are there features that predict the political party a state voted for in 2012? Is there a significant difference between female vs. male median incomes? What is the distribution of educational degrees and how does it correlate with income? How do race distributions vary between states? Is a state’s average temperature predictive of any feature? etc.

Each student will be asked to present his findings to the class.

**Requirements**

Each project should include at least one example:

* Data Visualization (histograms, box plots, scatter plots)
* Classification (decision tree or SVM; show accuracy and confusion mat)
* Linear regression (show R2)
* Clustering (kmeans or hierarchical)
* Statistical test (t-test)

**IPython Notebook**

Each student will turn in an IPython notebook that includes:

* Introduction (What did you do?)
* Data Analysis
* Conclusion (What did you learn?)

**Data Dictionary**

|  |  |
| --- | --- |
| **Field** | **Description** |
| State | State |
| City | City |
| TotalPopulation | Total city population |
| PercentMale | Percent male residents |
| PercentFemale | Percent female residents |
| MedianAge | Median age of residents in city |
| PercentOver65 | Percent population over 65 years |
| PercentWhite | Percent population of white descent |
| PercentBlack | Percent population of black descent |
| PercentHispanic | Percent population of hispanic descent |
| PercentAsian | Percent population of asian descent |
| MedianHouseholdIncome | Median household income |
| MeanHouseholdIncome | Mean household income |
| MedianIncomeFemale | Median income for females |
| MedianIncomeMale | Median income for males |
| AverageHouseholdSize | Average household size |
| AverageFamilySize | Average family size |
| EdElementaryPercent | % with elementary education |
| EdHighSchoolPercent | % with high school education |
| EdAssociatePercent | % with associates degree |
| EdBachelorsPercent | % with bachelor’s degree |
| EdGraduatePercent | % with graduate degree |
| AvgLowTemp | Average annual low temperature |
| AvgHighTemp | Average annual high temperature |
| ElectionResults | 2012 election results (red = republican, blue = democrat) |