**BIG PICTURE**

Problem statement:

* Classification vs. Numerical Prediction (model)
* Supervised Learning vs. Unsupervised Learning (algorithm)
  + Unsupervised: clustering (pre-processing tool) and association rules
    - Use clustering after choosing models

Pre-processing:

* Reduce columns
  + Correlation analysis
    - Some higher correlation between variables is acceptable (intuitively)
  + Formal statistical analysis (chi-squared test)
  + PCA (Principal Component Analysis)
    - Address correlation?
  + Scalability issue? Random sampling!

Normalization:

* Neural network
* Two popular normalization methods:
  + Standardization
  + Mean-max (most popular)
* Clustering
  + K-means algorithm (Euclidean distance from centroid)
* Discretization -> address in paper
  + Example: level/severity of crime
  + Categories?
* Numeric coding of nominal data
  + Convert to numerical in order to feed into model such as neural network

-----

Mutiple models

Clustering (3 clusters)

-----

Evaluation:

* Classification:
  + Confusion matrix
  + Ratios (accuracy, precision, recall, etc.)
* Numerical Prediction:
  + MSE

Contributions

Findings/Conclusions

-----

Domain knowledge – who is our audience?

* Something to highly consider throughout the process/paper
* New pattern or new knowledge? -> from our process/findings

**\*Week 7 expectation: discuss models\***

-----

Choosing a model/algorithm:

* Start with a decision tree
* Logit model? (odds-ratio)

-----

Normalization of Data

-----

Group Presentations:

* Appraisal Accuracy
  + Improve error rates of Automated Valuation Methods
    - But include community variables
  + Grouping: non-property, property, and violent crime summary
  + Correlation matrix: half triangle with heat scale
    - P-value of chi-square test
  + Model selection: decision tree and neural networks -> evaluate performance of models based on error rates (benchmarks)
* Different Crimes Affect Property Value
  + Residential property sold or above appraisal value, estimate impact different types of crime on property value
* Frequency of crime, school system characteristics, demographics, and housing characteristics
  + Revgeo package in R (not python)... (convert lat and long to zip codes)
* Influential features of residential property value
  + Data mining algorithms
  + Top 10 influential factors on residential property price in pierce county
  + Gis data sourse ->
  + Range, township, crime\_num

Group Meeting Week 6:

* Present at least one model
  + \*Decision Tree (lots of advantages)
    - Try random forest after constructing a decision tree
    - Example of R code on canvas
  + \*Neural Network (use at least 1 hidden layer; try 2-3 nodes)
    - 2 different parameter values
* Possibly 3 models
* Remember to keep up with the documentation of the project
  + Consolidate into final paper in shared drive