Data Recoding in Python

What is Recoding?

- Converting data from one format into another
- Sometimes only numeric data is required for a function
- You can pull additional meaning out of your data by re-grouping

Types of Recoding

- Same Variable
- New Variable
- Continuous to Categorical
- Dummy Coding
- Grouping with Recoding

New Variable

- Safest

- Create a recode function

```
def functionName (input):
        if input  == “oldValue”:
                return newValue


dataFrame[‘newColumn’] =
dataFrame[‘oldColumn’].apply(functionName)
```

Continuous to Categorical

- Use the same format as creating a new variable

- Groups numeric data by operators
  - >
  - <
  - >=
  - <=

Same Variable

- Use a dictionary of values and then replace them
- Can only be done once per dataset, so can cause more re-work
- Overwriting the current data

```
dictName = {“colName” : {“oldValue” : “newValue”}
dataFrame.replace(dictName, inplace=True)
```

Dummy Coding

- Technically the most correct for ML and Modeling situations

- Recodes every variable as binary, then add it back in

```
newDataFrame = pd.get_dummies(oldData['column'], drop_first=True)
newDataFrame2 = pd.concat([oldData, newDataFrame], axis = 1)
```