Kyle S. Speaker Notes:

Board 8

-visual captures the steps that were taken

-Libraries: pandas, numpy, scikitlearn, yellowbrick

-pyodbc (Python open database connectivity) interface to access data from database

-missing values and number showing up categorically (Year column)

-Ordinal Encoder (objects) and Standard Scaler (floats)

-train test split w/ features and output (vintage rating)

Board 9

-Feature selection – started w/ 11, some duplicate or irrelevant got down to 7 after dropping known useless and through some trial and error

Board 10

-Pearson Correlation Ranking of 5 features to understand the data’s correlation

-Estimating the model (R2 of 87-88)

-evaluating validity and usefulness (R2 of 86-87)

Board 11

-Similar R2 in the training and test sets (slightly better test) with an interesting distribution due to the ratings being made to the 10th and predictions spanning in between

-MSE low, overall a good model

Board 12

-start simple – NN was a lot of effort and time for terrible results

-while the model is a good fit, it’s usefulness is questionable as all the data are categorical or lag measure. It’d be more interesting and useful if we had more data about the wine (other than it being a vintage) like avg weekly sales or distribution spread etc. and predicted rankings that way.