import relevant libraries

Q/A:

**import** **statsmodels.api** **as** **sm** *# What does this do? -offers method in categories of Regression,*

*#Imputation, Generalized Estimating Equations, Generalized Linear Models,*

*#Discrete and Count Models, Multivariate Models, Misc Models, Graphics, Statistics, Tools*

**from** **statsmodels.graphics.api** **import** abline\_plot *# What does this do? -plots a line given an intercept and slope*

**from** **sklearn.metrics** **import** mean\_squared\_error, r2\_score *# What does this do? r2\_score gives the (cofficient of determination) regression score function*

*# mean\_squared\_error gives the mean squared error regression loss*

**from** **sklearn.model\_selection** **import** train\_test\_split *# What does this do? -splits arrays of matrices into random train and test subsets.*

**from** **sklearn** **import** linear\_model, preprocessing *# What does this do? -Linear\_model allows for LinearRegression()*

*# preprocessing package from sklearn offering utility functions and transformer*

*# classes to change raw feature vectors in to a form suitable for downstream estimators*

Q:

What can you infer about the nature of these variables, as output by the info() method?

Which variables might be suitable for regression analysis, and why? For those variables that aren't suitable for regression analysis, is there another type of statistical modeling for which they are suitable?

A:

all variables are numeric

all variables are suitable for regression analysis because they are all numeric

for variables that are not suitable for regression analysis a histogram per category is appropriate

Q:

We can see so much about the quality variable just from this simple visualization.

Answer yourself: what value do most wines have for quality?

What is the minimum quality value below, and the maximum quality value?

What is the range? Remind yourself of these summary statistical concepts by looking at p.49 of the AoS.

But can you think of a problem with making this variable the dependent variable of regression analysis?

Remember the example in AoS on p.122 of predicting the heights of children from the heights of parents?

Take a moment here to think about potential problems before reading on.

A:

Most wins have 5 for quality with 6 as a close second

The minimum quality wine value is 3. The maximum quality wine value is 8.

I cannot think of a problem with making this variable (i.e. quality) the dependent variable of regression analysis.

Yes, I remember the example of predicting the height of children from the heights of parents in AoS.

I do not see regression to the mean as a problem in regression.

\*\*\*quality is a discrete variable not a continuous variable here.

Q:

*# What do you notice from this summary?*

Out[85]:

count 1599.000000

mean 5.636023

std 0.807569

min 3.000000

25% 5.000000

50% 6.000000

75% 6.000000

max 8.000000

Name: quality, dtype: float64

A:

All quartiles are whole numbers

Q:

# Calling .value\_counts() on the quality field with the parameter dropna=False,

# get a list of the values of the quality variable, and the number of occurrences of each.

# **Do you know why we're calling value\_counts() with the parameter dropna=False?** Take a moment to research the

# answer if you're not sure.

A:

dropna=False is set inside .value\_count() because by default dropna=True and we want to keep rows with missing values for now.

Q:

If you've never executed your own Seaborn pairplot before, just take a moment to look at the output. They certainly output a lot of information at once. What can you infer from it? What can you not justifiably infer from it?

... All done?

A:

I can infer from the pairplot that some variables have more correlation with each other than others.

‘citric.acid’ and ‘fixed.acidity’ appear to have a positive correlation with ‘density’

‘fixed.acidity’ appears to have a positive correlation with density and a negative correlation with ‘pH’

‘total.sulfur.oxide’ and ‘free.sulfur.oxide’ appear to have a positive correlation

‘quality’ and ‘alcohol’ positive correlation

Negative correlations to ‘**alcohol**’ include ‘density’, ‘chloride’, ‘total.sulfur.dioxide’, and ‘volatile.acidity’

Positive correlations to ‘alcohol’ include ‘quality’ and ‘pH’

I cannot justifiably infer from the sns pairplot the numerical pearson correlation without calculation, though we can try to estimate positive or negative correlation by eacn cell.

Q:

Take a moment to think about the following questions:

* How does color relate to extent of correlation?
* How might we use the plot to show us interesting relationships worth investigating?
* More precisely, what does the heatmap show us about the fixed acidity variable's relationship to the density variable?

A:

The lighter the color on the heatmap the more positively correlated the variables. The darker the color on the heatmap the more negatively correlated the variables.

The very dark and very light colors on the heatmap indicated a strong negative and positive correlation respectively.

One or more of these strong correlations may be of interest in creating a linear regression.

The ‘fixed acidity’ variable has a strong positive correlation with the ‘density’ variable.