Springboard Regression Case Study, Unit 8 - The Red Wine Dataset - Tier 3

Welcome to the Unit 8 Springboard Regression case study! Please note: this is ***Tier 3*** of the case study.

This case study was designed for you to **use Python to apply the knowledge you've acquired in reading *The Art of Statistics* (hereinafter *AoS*) by Professor Spiegelhalter**. Specifically, the case study will get you doing regression analysis; a method discussed in Chapter 5 on p.121. It might be useful to have the book open at that page when doing the case study to remind you of what it is we're up to (but bear in mind that other statistical concepts, such as training and testing, will be applied, so you might have to glance at other chapters too).

The aim is to ***use exploratory data analysis (EDA) and regression to predict alcohol levels in wine with a model that's as accurate as possible***.

We'll try a *univariate* analysis (one involving a single explanatory variable) as well as a *multivariate* one (involving multiple explanatory variables), and we'll iterate together towards a decent model by the end of the notebook. The main thing is for you to see how regression analysis looks in Python and jupyter, and to get some practice implementing this analysis.

Throughout this case study, **questions** will be asked in the markdown cells. Try to **answer these yourself in a simple text file** when they come up. Most of the time, the answers will become clear as you progress through the notebook. Some of the answers may require a little research with Google and other basic resources available to every data scientist.

For this notebook, we're going to use the red wine dataset, wineQualityReds.csv. Make sure it's downloaded and sitting in your working directory. This is a very common dataset for practicing regression analysis and is actually freely available on Kaggle, here (https://www.kaggle.com/piyushgoyal443/red-wine-dataset).

You're pretty familiar with the data science pipeline at this point. This project will have the following structure: **1. Sourcing and loading**

- Import relevant libraries
- Load the data
- Exploring the data
- Choosing a dependent variable

2. Cleaning, transforming, and visualizing

- Visualizing correlations

3. Modeling

- Train/Test split
- Making a Linear regression model: your first model
- Making a Linear regression model: your second model: Ordinary Least Squares (OLS)
- Making a Linear regression model: your third model: multiple linear regression

- Making a Linear regression model: your fourth model: avoiding redundancy

4. Evaluating and concluding

- Reflection
- Which model was best?
- Other regression algorithms

1. Sourcing and loading

1a. Import relevant libraries

In [3]:
```
# Import relevant libraries and packages.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns # For all our visualization needs.
import statsmodels.api as sm # What does this do? Find out and type here.
from statsmodels.graphics.api import abline_plot # What does this do? Find
from sklearn.metrics import mean_squared_error, r2_score # What does this d
from sklearn.model_selection import train_test_split #  What does this do?
from sklearn import linear_model, preprocessing # What does this do? Find o
import warnings # For handling error messages.
# Don't worry about the following two instructions: they just suppress warn
warnings.simplefilter(action="ignore", category=FutureWarning)
warnings.filterwarnings(action="ignore", module="scipy", message="^internal
```

1b. Load the data

In [22]:
```
# Load the data.
wine_data = pd.DataFrame(pd.read_csv("winequalityreds.csv"))
wine_data = wine_data.drop(columns = ['Unnamed: 0'])
```

1c. Exploring the data

In [23]:
```
# Check out its appearance.
wine_data.head()
```

Out[23]:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dic
0	7.4	0.70	0.00	1.9	0.076	11.0	
1	7.8	0.88	0.00	2.6	0.098	25.0	
2	7.8	0.76	0.04	2.3	0.092	15.0	
3	11.2	0.28	0.56	1.9	0.075	17.0	
4	7.4	0.70	0.00	1.9	0.076	11.0	

In [24]:
```
# Another very useful method to call on a recently imported dataset is .inf
# overview of the data
wine_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed.acidity         1599 non-null   float64
 1   volatile.acidity      1599 non-null   float64
 2   citric.acid           1599 non-null   float64
 3   residual.sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free.sulfur.dioxide   1599 non-null   float64
 6   total.sulfur.dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

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?

In [25]:
```
# We should also look more closely at the dimensions of the dataset.
wine_data.shape
```

Out[25]:
```
(1599, 12)
```

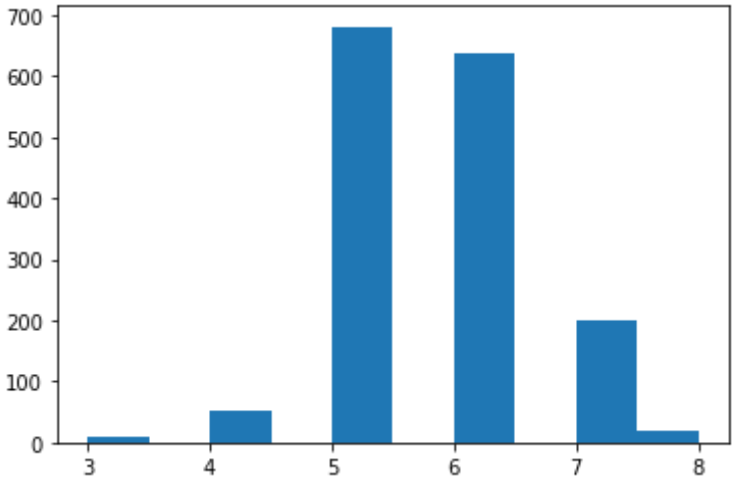
1d. Choosing a dependent variable

We now need to pick a dependent variable for our regression analysis: a variable whose values we will predict.

'Quality' seems to be as good a candidate as any. Let's check it out. One of the quickest and most informative ways to understand a variable is to make a histogram of it. This gives us an idea of both the center and spread of its values.

In [37]:
```
# Making a histogram of the quality variable.
plt.hist(wine_data[['quality']])
```

Out[37]:
```
(array([ 10.,   0.,  53.,   0., 681.,   0., 638.,   0., 199.,  18.]),
 array([3. , 3.5, 4. , 4.5, 5. , 5.5, 6. , 6.5, 7. , 7.5, 8. ]),
 <BarContainer object of 10 artists>)
```



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.

The issue is this: quality is a *discrete* variable, in that its values are integers (whole numbers) rather than floating point numbers. Thus, quality is not a *continuous* variable. But this means that it's actually not the best target for regression analysis.

Before we dismiss the quality variable, however, let's verify that it is indeed a discrete variable with some further exploration.

In [27]:
```
# Get a basic statistical summary of the variable
wine_data['quality'].describe()
```

Out[27]:
```
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
```

In [34]:
```
# Get a list of the values of the quality variable, and the number of occur
wine_data['quality'].value_counts(dropna=False)
```

Out[34]:
```
5    681
6    638
7    199
4     53
8     18
3     10
Name: quality, dtype: int64
```

The outputs of the describe() and value_counts() methods are consistent with our histogram, and since there are just as many values as there are rows in the dataset, we can infer that there are no NAs for the quality variable.

But scroll up again to when we called info() on our wine dataset. We could have seen there, already, that the quality variable had int64 as its type. As a result, we had sufficient information, already, to know that the quality variable was not appropriate for regression analysis. Did you figure this out yourself? If so, kudos to you!

The quality variable would, however, conduce to proper classification analysis. This is because, while the values for the quality variable are numeric, those numeric discrete values represent *categories*; and the prediction of category-placement is most often best done by classification algorithms. You saw the decision tree output by running a classification algorithm on the Titanic dataset on p.168 of Chapter 6 of *AoS*. For now, we'll continue with our regression analysis, and continue our search for a suitable dependent variable.

Now, since the rest of the variables of our wine dataset are continuous, we could — in theory — pick any of them. But that does not mean that are all equally sutiable choices. What counts as a suitable dependent variable for regression analysis is determined not just by *intrinsic* features of the dataset (such as data types, number of NAs etc) but by *extrinsic* features, such as, simply, which variables are the most interesting or useful to predict, given our aims and values in the context we're in. Almost always, we can only determine which variables are sensible choices for dependent variables with some **domain knowledge**.

Not all of you might be wine buffs, but one very important and interesting quality in wine is [acidity (https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity)](https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity). As the Waterhouse Lab at the University of California explains, 'acids impart the sourness or tartness that is a fundamental

feature in wine taste. Wines lacking in acid are "flat." Chemically the acids influence titrable acidity which affects taste and pH which affects color, stability to oxidation, and consequantly the overall lifespan of a wine.'

If we cannot predict quality, then it seems like **fixed acidity** might be a great option for a dependent variable. Let's go for that.

So if we're going for fixed acidity as our dependent variable, what we now want to get is an idea of *which variables are related interestingly to that dependent variable*.

We can call the .corr() method on our wine data to look at all the correlations between our variables. As the documentation (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html) shows, the default correlation coefficient is the Pearson correlation coefficient (p.58 and p.396 of the *AoS*); but other coefficients can be plugged in as parameters. Remember, the Pearson correlation coefficient shows us how close to a straight line the data-points fall, and is a number between -1 and 1.

In [29]:
```
# Call the .corr() method on the wine dataset
wine_data.corr()
```

Out[29]:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxic
fixed.acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.15379
volatile.acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.01050
citric.acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.06097
residual.sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.18704
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.00556
free.sulfur.dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.00000
total.sulfur.dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.66766
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.02194
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.07037
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.05165
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.06940
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.05065

Ok - you might be thinking, but wouldn't it be nice if we visualized these relationships? It's hard to get a picture of the correlations between the variables without anything visual.

Very true, and this brings us to the next section.

2. Cleaning, Transforming, and Visualizing

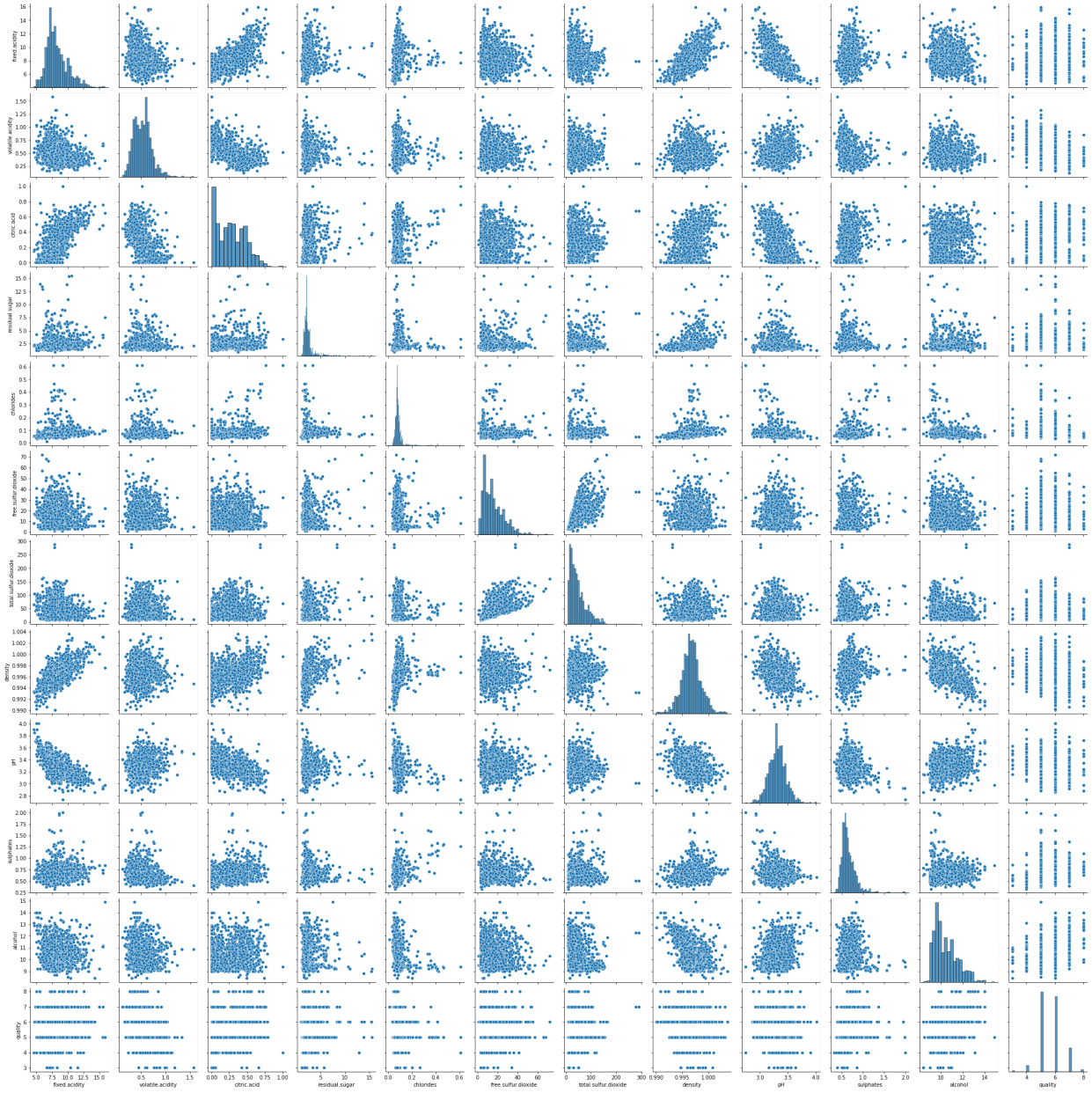
2a. Visualizing correlations

The heading of this stage of the data science pipeline ('Cleaning, Transforming, and Visualizing') doesn't imply that we have to do all of those operations in *that order*. Sometimes (and this is a case in point) our data is already relatively clean, and the priority is to do some visualization. Normally, however, our data is less sterile, and we have to do some cleaning and transforming first prior to visualizing.

Now that we've chosen alcohol level as our dependent variable for regression analysis, we can begin by plotting the pairwise relationships in the dataset, to check out how our variables relate to one another.

In [30]:
```
# Make a pairplot of the wine data
sns.pairplot(wine_data)
```

Out[30]:
```
<seaborn.axisgrid.PairGrid at 0x1a304aef90>
```



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?

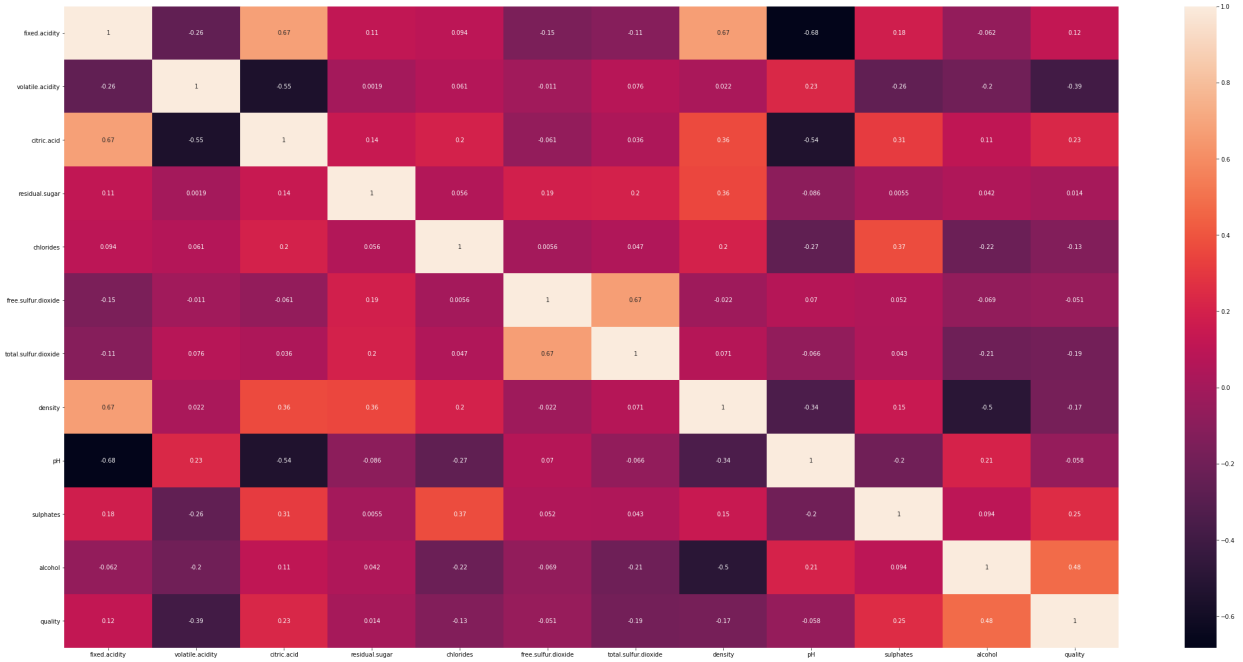
Here's a couple things you might have noticed:

- a given cell value represents the correlation that exists between two variables
- on the diagonal, you can see a bunch of histograms. This is because pairplotting the variables with themselves would be pointless, so the pairplot() method instead makes histograms to show the distributions of those variables' values. This allows us to quickly see the shape of each variable's values.
- the plots for the quality variable form horizontal bands, due to the fact that it's a discrete variable. We were certainly right in not pursuing a regression analysis of this variable.
- Notice that some of the nice plots invite a line of best fit, such as alcohol vs density. Others, such as citric acid vs alcohol, are more inscrutable.

So we now have called the .corr() method, and the .pairplot() Seaborn method, on our wine data. Both have flaws. Happily, we can get the best of both worlds with a heatmap.

In [36]:
```
# Make a heatmap of the data
plt.figure(figsize=(40, 20))
sns.heatmap(wine_data.corr(), annot=True)
```

Out[36]: <AxesSubplot:>



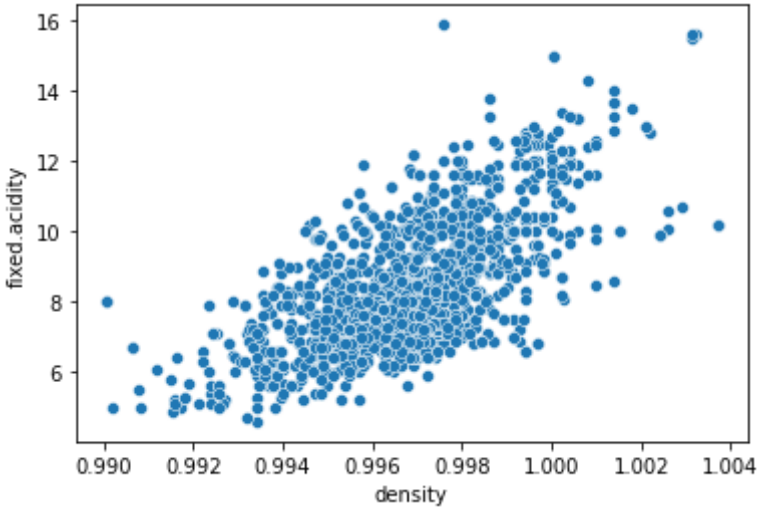
Take a moment to think about the following questions:

- 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?

There is a relatively strong correlation between the density and fixed acidity variables respectively. In the next code block, call the scatterplot() method on our sns object. Make the x-axis parameter 'density', the y-axis parameter 'fixed.acidity', and the third parameter specify our wine dataset.

In [32]:
```
# Plot density against alcohol
sns.scatterplot(data=wine_data, x='density', y='fixed.acidity')
```

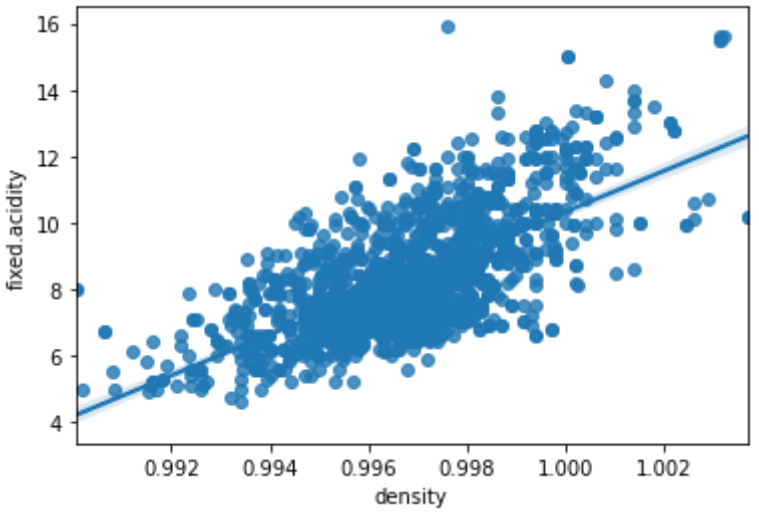
Out[32]:
```
<AxesSubplot:xlabel='density', ylabel='fixed.acidity'>
```



We can see a positive correlation, and quite a steep one. There are some outliers, but as a whole, there is a steep looking line that looks like it ought to be drawn.

In [33]:
```
# Call the regplot method on your sns object, with parameters: x = 'density
sns.regplot(data=wine_data, x='density', y='fixed.acidity')
```

Out[33]:
```
<AxesSubplot:xlabel='density', ylabel='fixed.acidity'>
```



The line of best fit matches the overall shape of the data, but it's clear that there are some points that deviate from the line, rather than all clustering close.

Let's see if we can predict fixed acidity based on density using linear regression.

3. Modeling

3a. Train/Test Split

While this dataset is super clean, and hence doesn't require much for analysis, we still need to split our dataset into a test set and a training set.

You'll recall from p.158 of *AoS* that such a split is important good practice when evaluating statistical models. On p.158, Professor Spiegelhalter was evaluating a classification tree, but the same applies when we're doing regression. Normally, we train with 75% of the data and test on the remaining 25%.

To be sure, for our first model, we're only going to focus on two variables: fixed acidity as our dependent variable, and density as our sole independent predictor variable.

We'll be using sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) here. Don't worry if not all of the syntax makes sense; just follow the rationale for what we're doing.

In [39]:
```
# Subsetting our data into our dependent and independent variables.
X = wine_data[['density']]
Y = wine_data[['fixed.acidity']]

# Split the data. This line uses the sklearn function train_test_split().
# The test_size parameter means we can train with 75% of the data, and test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,
```

In [42]:
```
# We now want to check the shape of the X train, y_train, X_test and y_test
X_train.shape, Y_train.shape
```

Out[42]:
```
((1199, 1), (1199, 1))
```

In [43]:
```
X_test.shape, Y_test.shape
```

Out[43]:
```
((400, 1), (400, 1))
```

3b. Making a Linear Regression model: our first model

Sklearn has a LinearRegression() (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) function built into the linear_model module. We'll be using that to make our regression model.

In [49]:
```
# Create the model: make a variable called rModel, and use it linear_model.
rModel = linear_model.LinearRegression(normalize=True)
```

In [50]:
```
# We now want to train the model on our test data.
rModel.fit(X_train, Y_train)
```

Out[50]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=
True)
```

In [51]:
```
# Evaluate the model
print(rModel.score(X_train, Y_train))
```

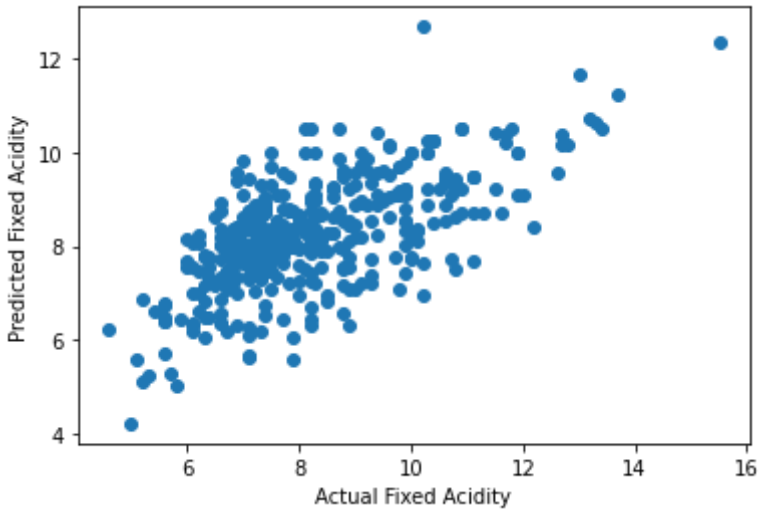
```
0.45487824100681673
```

The above score is called R-Squared coefficient, or the "coefficient of determination". It's basically a measure of how successfully our model predicts the variations in the data away from the mean: 1 would mean a perfect model that explains 100% of the variation. At the moment, our model explains only about 23% of the variation from the mean. There's more work to do!

In [52]:
```
# Use the model to make predictions about our test data
ypred = rModel.predict(X_test)
```

In [55]:
```
# Let's plot the predictions against the actual result. Use scatter()
plt.scatter(Y_test, ypred)
plt.xlabel('Actual Fixed Acidity')
plt.ylabel('Predicted Fixed Acidity')
```

Out[55]:
```
Text(0, 0.5, 'Predicted Fixed Acidity')
```



The above scatterplot represents how well the predictions match the actual results.

Along the x-axis, we have the actual fixed acidity, and along the y-axis we have the predicted value for the fixed acidity.

There is a visible positive correlation, as the model has not been totally unsuccesful, but it's clear that it is not maximally accurate: wines with an actual fixed acidity of just over 10 have been predicted as having acidity levels from about 6.3 to 13.

Let's build a similar model using a different package, to see if we get a better result that way.

3c. Making a Linear Regression model: our second model: Ordinary Least Squares (OLS)

In [73]:
```
# Create the test and train sets. Here, we do things slightly differently.
# We make the explanatory variable X as before.
X = wine_data[['density']]

# But here, reassign X the value of adding a constant to it. This is requir
# Further explanation of this can be found here:
# https://www.statsmodels.org/devel/generated/statsmodels.regression.linear

X = sm.add_constant(X, prepend=True)
print(X)
```

```
      const  density
0       1.0  0.99780
1       1.0  0.99680
2       1.0  0.99700
3       1.0  0.99800
4       1.0  0.99780
...     ...      ...
1594    1.0  0.99490
1595    1.0  0.99512
1596    1.0  0.99574
1597    1.0  0.99547
1598    1.0  0.99549

[1599 rows x 2 columns]
```

In [74]:
```
# The rest of the preparation is as before.
Y = wine_data[['fixed.acidity']]

# Split the data using train_test_split()
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,
```

In [75]:
```
X_train.shape, Y_train.shape
```

Out[75]:
```
((1199, 2), (1199, 1))
```

In [76]:
```
X_test.shape, Y_test.shape
```

Out[76]:
```
((400, 2), (400, 1))
```

In [77]:
```
# Create the model
rModel2 = sm.OLS(Y_train, X_train)

# Fit the model with fit()
rModel2results = rModel2.fit()
```

In [78]:
```
# Evaluate the model with .summary()
rModel2results.summary()
```

Out[78]:

OLS Regression Results

Dep. Variable:	fixed.acidity	**R-squared:**	0.455
Model:	OLS	**Adj. R-squared:**	0.454
Method:	Least Squares	**F-statistic:**	998.8
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	6.68e-160
Time:	13:25:01	**Log-Likelihood:**	-2011.0
No. Observations:	1199	**AIC:**	4026.
Df Residuals:	1197	**BIC:**	4036.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-615.7316	19.746	-31.182	0.000	-654.473	-576.990
density	626.0927	19.810	31.604	0.000	587.226	664.959

Omnibus:	94.056	**Durbin-Watson:**	1.985
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	122.229
Skew:	0.668	**Prob(JB):**	2.87e-27
Kurtosis:	3.812	**Cond. No.**	1.06e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.06e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

One of the great things about Statsmodels (sm) is that you get so much information from the summary() method.

There are lots of values here, whose meanings you can explore at your leisure, but here's one of the most important: the R-squared score is 0.455, the same as what it was with the previous model. This makes perfect sense, right? It's the same value as the score from sklearn, because they've both used the same algorithm on the same data.

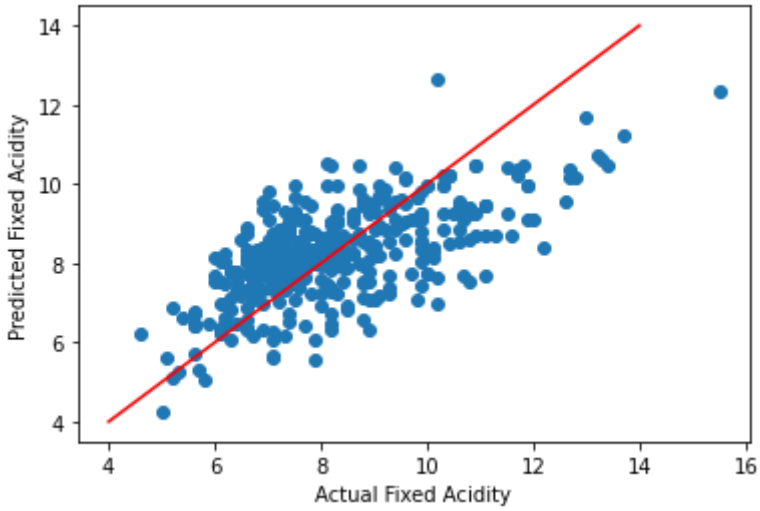
Here's a useful link you can check out if you have the time:
https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
(https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/)

In [79]:
```
# Let's use our new model to make predictions of the dependent variable y.
ypred2 = rModel2results.predict(X_test)
```

In [83]:
```
# Plot the predictions
# Build a scatterplot
plt.scatter(Y_test, ypred2)
# Add a line for perfect correlation. Can you see what this line is doing?
plt.plot([x for x in range(4,15)],[x for x in range(4,15)], color='red')

# Label it nicely
plt.xlabel('Actual Fixed Acidity')
plt.ylabel('Predicted Fixed Acidity')
```

Out[83]:
```
Text(0, 0.5, 'Predicted Fixed Acidity')
```



The red line shows a theoretically perfect correlation between our actual and predicted values - the line that would exist if every prediction was completely correct. It's clear that while our points have a generally similar direction, they don't match the red line at all; we still have more work to do.

To get a better predictive model, we should use more than one variable.

3d. Making a Linear Regression model: our third model: multiple linear regression

Remember, as Professor Spiegelhalter explains on p.132 of *AoS*, including more than one explanatory variable into a linear regression analysis is known as ***multiple linear regression***.

In [86]:
```
# Create test and train datasets
# This is again very similar, but now we include more columns in the predic
# Include all columns from data in the explanatory variables X except fixed
X = wine_data.drop(columns={'fixed.acidity', 'quality'}, axis=1)

# Create constants for X, so the model knows its bounds
X = sm.add_constant(X)
y = wine_data[['fixed.acidity']]

# Split the data
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,
```

In [87]:
```
# We can use almost identical code to create the third model, because it is
# Create the model
rModel3 = sm.OLS(Y_train, X_train)

# Fit the model
rModel3results = rModel3.fit()
```

In [89]:
```
# Evaluate the model
rModel3results.summary()
```

Out[89]:

OLS Regression Results

Dep. Variable:	fixed.acidity	**R-squared:**	0.871
Model:	OLS	**Adj. R-squared:**	0.870
Method:	Least Squares	**F-statistic:**	804.4
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	0.00
Time:	13:36:54	**Log-Likelihood:**	-1145.6
No. Observations:	1199	**AIC:**	2313.
Df Residuals:	1188	**BIC:**	2369.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-648.2418	15.246	-42.518	0.000	-678.154	-618.329
volatile.acidity	0.1313	0.135	0.971	0.332	-0.134	0.397
citric.acid	1.8651	0.155	11.995	0.000	1.560	2.170
residual.sugar	-0.2485	0.015	-16.589	0.000	-0.278	-0.219
chlorides	-3.6575	0.443	-8.263	0.000	-4.526	-2.789
free.sulfur.dioxide	0.0068	0.002	2.859	0.004	0.002	0.012
total.sulfur.dioxide	-0.0064	0.001	-7.966	0.000	-0.008	-0.005
density	671.0968	15.184	44.198	0.000	641.306	700.887
pH	-5.1954	0.149	-34.792	0.000	-5.488	-4.902
sulphates	-0.8038	0.130	-6.193	0.000	-1.058	-0.549
alcohol	0.5713	0.025	22.753	0.000	0.522	0.621

Omnibus:	155.238	**Durbin-Watson:**	2.052
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	562.315
Skew:	0.595	**Prob(JB):**	7.85e-123
Kurtosis:	6.137	**Cond. No.**	7.23e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.23e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

The R-Squared score shows a big improvement - our first model predicted only around 45% of the variation, but now we are predicting 87%!

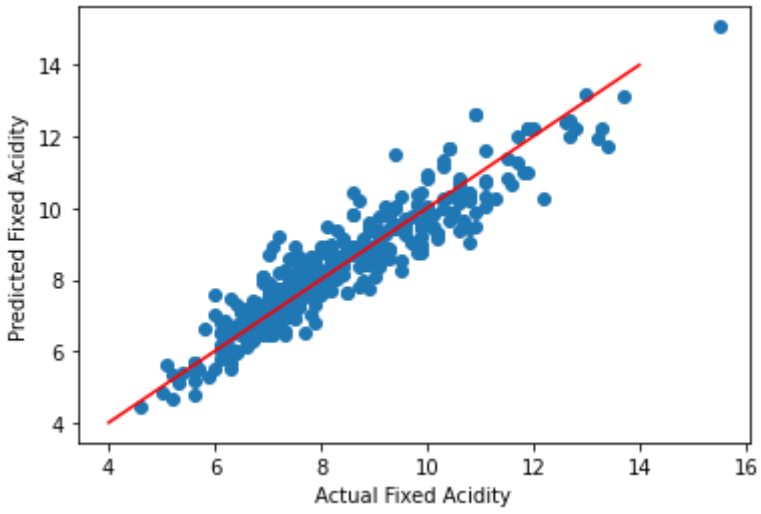
In [91]:
```
# Use our new model to make predictions
ypred3 = rModel3results.predict(X_test)
```

In [92]:
```
# Plot the predictions
# Build a scatterplot
plt.scatter(Y_test, ypred3)

# Add a line for perfect correlation
plt.plot([x for x in range(4,15)],[x for x in range(4,15)], color='red')

# Label it nicely
plt.xlabel('Actual Fixed Acidity')
plt.ylabel('Predicted Fixed Acidity')
```

Out[92]:
```
Text(0, 0.5, 'Predicted Fixed Acidity')
```



We've now got a much closer match between our data and our predictions, and we can see that the shape of the data points is much more similar to the red line.

We can check another metric as well - the RMSE (Root Mean Squared Error). The MSE is defined by Professor Spiegelhalter on p.393 of *AoS*, and the RMSE is just the square root of that value. This is a measure of the accuracy of a regression model. Very simply put, it's formed by finding the average difference between predictions and actual values. Check out p. 163 of *AoS* for a reminder of how this works.

In [93]:
```
# Define a function to check the RMSE. Remember the def keyword needed to m
def findRMSE(predictions, targets):
    return np.sqrt(((predictions - targets) ** 2).mean())
```

In [95]:
```
# Get predictions from rModel3
ypred3 = rModel3results.predict(X_test)

# Put the predictions & actual values into a dataframe
matches = pd.DataFrame(Y_test)
matches.rename(columns = {'fixed.acidity':'actual'}, inplace=True)
matches["predicted"] = ypred3

findRMSE(matches["actual"], matches["predicted"])
```

Out[95]:
```
0.6163194678948996
```

The RMSE tells us how far, on average, our predictions were mistaken. An RMSE of 0 would mean we were making perfect predictions. 0.6 signifies that we are, on average, about 0.6 of a unit of fixed acidity away from the correct answer. That's not bad at all.

3e. Making a Linear Regression model: our fourth model: avoiding redundancy

We can also see from our early heat map that volatile.acidity and citric.acid are both correlated with pH. We can make a model that ignores those two variables and just uses pH, in an attempt to remove redundancy from our model.

In [98]:
```
# Create test and train datasets
X = wine_data[["residual.sugar","chlorides","total.sulfur.dioxide","density

# Create constants for X, so the model knows its bounds
X = sm.add_constant(X)
Y = wine_data[["fixed.acidity"]]

# Split the data

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25,
```

In [99]:
```
# Create the fifth model
rModel4 = sm.OLS(Y_train, X_train)
# Fit the model
rModel4results = rModel4.fit()
# Evaluate the model
rModel4results.summary()
```

Out[99]:

OLS Regression Results

Dep. Variable:	fixed.acidity	**R-squared:**	0.742
Model:	OLS	**Adj. R-squared:**	0.741
Method:	Least Squares	**F-statistic:**	571.8
Date:	Sat, 17 Oct 2020	**Prob (F-statistic):**	0.00
Time:	14:16:16	**Log-Likelihood:**	-1562.3
No. Observations:	1199	**AIC:**	3139.
Df Residuals:	1192	**BIC:**	3174.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-485.6576	16.010	-30.335	0.000	-517.068	-454.247
residual.sugar	-0.1078	0.020	-5.481	0.000	-0.146	-0.069
chlorides	-6.3544	0.578	-10.990	0.000	-7.489	-5.220
total.sulfur.dioxide	-0.0094	0.001	-11.799	0.000	-0.011	-0.008
density	516.4441	15.894	32.492	0.000	485.260	547.628
pH	-6.0430	0.184	-32.766	0.000	-6.405	-5.681
sulphates	0.7540	0.165	4.559	0.000	0.430	1.078

Omnibus:	105.987	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	206.837
Skew:	0.572	**Prob(JB):**	1.22e-45
Kurtosis:	4.683	**Cond. No.**	5.08e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.08e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

The R-squared score has reduced, showing us that actually, the removed columns were important.

Conclusions & next steps

Congratulations on getting through this implementation of regression and good data science practice in Python!

Take a moment to reflect on which model was the best, before reading on.

. . .

Here's one conclusion that seems right. While our most predictively powerful model was rModel3, this model had explanatory variables that were correlated with one another, which made some redundancy. Our most elegant and economical model was rModel4 - it used just a few predictors to get a good result.

All of our models in this notebook have used the OLS algorithm - Ordinary Least Squares. There are many other regression algorithms, and if you have time, it would be good to investigate them. You can find some examples here (https://www.statsmodels.org/dev/examples/index.html#regression). Be sure to make a note of what you find, and chat through it with your mentor at your next call.