appending

October 18, 2017

1 Appending Data

First, import the necessary packages and load `winequality-red.csv` and `winequality-white.csv`.

```
In [2]: # import numpy and pandas
        import pandas as pd
        import numpy as np

        # load red and white wine datasets
        red_df = pd.read_csv('winequality-red.csv', sep=';')
        white_df = pd.read_csv('winequality-white.csv', sep=';')
```

1.1 Create Color Columns

Create two arrays as long as the number of rows in the red and white dataframes that repeat the value "red" or "white." NumPy offers really easy way to do this. Here's the documentation for NumPy's repeat function. Take a look and try it yourself.

```
In [7]: # create color array for red dataframe
        color_red = np.repeat('RED', red_df.shape[0])

        # create color array for white dataframe
        color_white = np.repeat('WHITE', white_df.shape[0])
```

Add arrays to the red and white dataframes. Do this by setting a new column called 'color' to the appropriate array. The cell below does this for the red dataframe.

```
In [8]: red_df['color'] = color_red
        red_df.head()
```

```
Out[8]:    fixed_acidity  volatile_acidity  citric_acid  residual_sugar  chlorides  \
        0            7.4              0.70         0.00             1.9      0.076
        1            7.8              0.88         0.00             2.6      0.098
        2            7.8              0.76         0.04             2.3      0.092
        3           11.2              0.28         0.56             1.9      0.075
        4            7.4              0.70         0.00             1.9      0.076
```

```
   free_sulfur_dioxide  total_sulfur-dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality color
0      9.4        5   RED
1      9.8        5   RED
2      9.8        5   RED
3      9.8        6   RED
4      9.4        5   RED
```

Do the same for the white dataframe and use `head()` to confirm the change.

```
In [9]: white_df['color'] = color_white
        white_df.head()
```

```
Out[9]:    fixed_acidity  volatile_acidity  citric_acid  residual_sugar  chlorides  \
        0            7.0              0.27         0.36            20.7      0.045
        1            6.3              0.30         0.34             1.6      0.049
        2            8.1              0.28         0.40             6.9      0.050
        3            7.2              0.23         0.32             8.5      0.058
        4            7.2              0.23         0.32             8.5      0.058

           free_sulfur_dioxide  total_sulfur_dioxide  density    pH  sulphates  \
        0                 45.0                 170.0   1.0010  3.00       0.45
        1                 14.0                 132.0   0.9940  3.30       0.49
        2                 30.0                  97.0   0.9951  3.26       0.44
        3                 47.0                 186.0   0.9956  3.19       0.40
        4                 47.0                 186.0   0.9956  3.19       0.40

           alcohol  quality  color
        0      8.8        6  WHITE
        1      9.5        6  WHITE
        2     10.1        6  WHITE
        3      9.9        6  WHITE
        4      9.9        6  WHITE
```

1.2 Combine DataFrames with Append

Check the documentation for Pandas' append function and see if you can use this to figure out how to combine the dataframes. (Bonus: Why aren't we using the merge method to combine the dataframes?) If you don't get it, I'll show you how afterwards. Make sure to save your work in this notebook! You'll come back to this later.

```
In [13]: # append dataframes
         wine_df = red_df.append(white_df)
```

```
# view dataframe to check for success
wine_df.head()
```

Out[13]:

	alcohol	chlorides	citric_acid	color	density	fixed_acidity	\
0	9.4	0.076	0.00	RED	0.9978	7.4	
1	9.8	0.098	0.00	RED	0.9968	7.8	
2	9.8	0.092	0.04	RED	0.9970	7.8	
3	9.8	0.075	0.56	RED	0.9980	11.2	
4	9.4	0.076	0.00	RED	0.9978	7.4	

	free_sulfur_dioxide	pH	quality	residual_sugar	sulphates	\
0	11.0	3.51	5	1.9	0.56	
1	25.0	3.20	5	2.6	0.68	
2	15.0	3.26	5	2.3	0.65	
3	17.0	3.16	6	1.9	0.58	
4	11.0	3.51	5	1.9	0.56	

	total_sulfur-dioxide	total_sulfur_dioxide	volatile_acidity
0	34.0	NaN	0.70
1	67.0	NaN	0.88
2	54.0	NaN	0.76
3	60.0	NaN	0.28
4	34.0	NaN	0.70

In [15]:
```
wine_df.tail()
```

Out[15]:

	alcohol	chlorides	citric_acid	color	density	fixed_acidity	\
4893	11.2	0.039	0.29	WHITE	0.99114	6.2	
4894	9.6	0.047	0.36	WHITE	0.99490	6.6	
4895	9.4	0.041	0.19	WHITE	0.99254	6.5	
4896	12.8	0.022	0.30	WHITE	0.98869	5.5	
4897	11.8	0.020	0.38	WHITE	0.98941	6.0	

	free_sulfur_dioxide	pH	quality	residual_sugar	sulphates	\
4893	24.0	3.27	6	1.6	0.50	
4894	57.0	3.15	5	8.0	0.46	
4895	30.0	2.99	6	1.2	0.46	
4896	20.0	3.34	7	1.1	0.38	
4897	22.0	3.26	6	0.8	0.32	

	total_sulfur-dioxide	total_sulfur_dioxide	volatile_acidity
4893	NaN	92.0	0.21
4894	NaN	168.0	0.32
4895	NaN	111.0	0.24
4896	NaN	110.0	0.29
4897	NaN	98.0	0.21

1.3 Save Combined Dataset

Save your newly combined dataframe as `winequality_edited.csv`. Remember, set `index=False` to avoid saving with an unnamed column!

```
In [16]: wine_df.to_csv('winequality_edited.csv', index=False)
```

```
In [ ]:
```