Day 9: *Data Filtering in Pandas*

After sightseeing in the Parisian streets, Dot boarded another plane for a short flight over to Florence. They barely had a chance to gaze out the window and eat a snack before landing in the great Italian city of art, culture, and fashion. In preparation to visit the Tuscan city that was the birthplace of modern Italian, Dot learned a few words of the language before landing. Grazie, they said to the flight attendant as they left the plane, buona giornata! The flight attendant smiled at Dot and waved them goodbye: arrivederci!

Dot made haste zipping from the airport straight to the art galleries Florence is so known for. They wanted to take in the work of the Renaissance masters that had shaped art history, and there's no better place for that than Florence. Dot spent the day hopping through the Uffizi gallery, the Palazzo Pitti, and the Accademia gallery. They even got a chance to look at Michaelangelo's famous statue of David! At the end of a busy day, Dot settled in at the patio of a swanky bar, hoping to get a taste of Tuscany's world-renowned wine. The waiter came and passed a booklet over to Dot, inviting them to open it up and read. Inside was a ton of information about the wines produced in Italy. Dot's eager to learn – let's help them examine this data!

Tutorial

Data in this tutorial are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.The analysis determined the quantities of 13 constituents found in each of the three types of wines.

In [18]:
```
import pandas as pd
df = pd.read_csv('wine.csv')
```

In [19]:
```
df.head()
```

Out[19]:

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyan
0	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	
1	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	
2	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	
3	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	
4	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	

An important concept in data analysis that you should become familiar with is *data filtering*. We can think about filtering as a two step process. The first step is creating a **boolean condition** that works as a filter, and the second step is passing the data through the filter.

```
#Example - Loading Data
user_filter = pd.read_csv('wine.csv')
df.head()

#Step 1: Create filter
user_filter = df['Alcohol'] >= 14

#Step 2: Feed the filter to the original DataFrame and store the result in a
new variable
filtered_df = df[user_filter]

#Step 3: Display Variable
filtered_df
```

Try the code above and see how it plays out. When you create a filter, observe the change in the number of rows between the old filtered DataFrame and the new one.

We can combine step 1 and 2 into one step:

```
# Step 1 and 2
filtered_df_2 = df[df['Alcohol'] >= 14]
```

Compare the two filtered DataFrames to see that they are the same. We can use the Python function **len()** to see how many rows the DataFrame has.

```
print(len(user_filter))
print(len(filtered_df))
print(len(filtered_df) == len(filtered_df_2))
```

To learn more about filtering in pandas, read this article.

To learn more about the various pandas functions, check out the user guide in the pandas documentation.

In [22]:
```
filtered_df_2 = df[df['Alcohol'] < 13]
filtered_df_2.head()
```

Challenge

Answer the following questions using the data:

- **How many Italian wines have lower percentage of alcohol than 13%**
- **How many wines are there in class 3?**

In [2]:
```
import pandas as pd
df = pd.read_csv('wine.csv')
df.head()
```

Out[2]:

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyan
0	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyan
1	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	
2	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	
3	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	
4	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	

Stretch Questions

- How many wines have level of magnesium between 90 and 100?
- How many wines have level of magnesium higher than 90, and a percentage of alcohol lower than 13.5%?

In [3]:
```
# SOLUTION

print(len(df[df.Alcohol < 13]))
print(len(df[df.Class == 3]))
```

```
86
48
```

In [4]:
```
# STRETCH SOLUTION

print(len(df[df.Magnesium.between(90,100)]))
print(len(df[(df.Magnesium > 90) & (df.Alcohol < 13.5)]))
```

```
48
71
```