Statistical Analysis:

Italian Wines

Melanie Straub

About this Dataset

- Chemical analysis of wines made by three cultivators in Italy
- Used to determine the origin of wines
- 13 Variables (Alcohol %, Malic Acid, Ash, Alkalinity, Magnesium, Phenols, Flavonoids, Nonflavonoid Phenols, Proanthocyanins, Color, Hue, OD280/OD315 of diluted wines, Proline)
- 178 Instances
- All attributes are continuous

Hypothesis

Research Hypothesis

Chemical constituents found in wine can help researches determine its origin by comparing its properties to other wines

Statistical Hypothesis

Alkalinity can be used to determine wine's origin due to the soil, climate and annual rainfall in each location

A look at the Dataset

```
print(wine)
```

```
     Wine  Alcohol  Malic.acid   Ash   Acl   Mg  Phenols  Flavanoids  \
0       1    14.23        1.71  2.43  15.6  127     2.80        3.06
1       1    13.20        1.78  2.14  11.2  100     2.65        2.76
2       1    13.16        2.36  2.67  18.6  101     2.80        3.24
3       1    14.37        1.95  2.50  16.8  113     3.85        3.49
4       1    13.24        2.59  2.87  21.0  118     2.80        2.69
..    ...      ...         ...   ...   ...  ...      ...         ...
173     3    13.71        5.65  2.45  20.5   95     1.68        0.61
174     3    13.40        3.91  2.48  23.0  102     1.80        0.75
175     3    13.27        4.28  2.26  20.0  120     1.59        0.69
176     3    13.17        2.59  2.37  20.0  120     1.65        0.68
177     3    14.13        4.10  2.74  24.5   96     2.05        0.76

     Nonflavanoid.phenols  Proanth  Color.int   Hue    OD  Proline
0                    0.28     2.29       5.64  1.04  3.92     1065
1                    0.26     1.28       4.38  1.05  3.40     1050
2                    0.30     2.81       5.68  1.03  3.17     1185
3                    0.24     2.18       7.80  0.86  3.45     1480
4                    0.39     1.82       4.32  1.04  2.93      735
..                    ...      ...        ...   ...   ...      ...
173                  0.52     1.06       7.70  0.64  1.74      740
174                  0.43     1.41       7.30  0.70  1.56      750
175                  0.43     1.35      10.20  0.59  1.56      835
176                  0.53     1.46       9.30  0.60  1.62      840
177                  0.56     1.35       9.20  0.61  1.60      560
```

Pandas Dataframe Correlation

- Used to show pairwise correlation for all column and row variables in the dataset
  - "1" shows perfect correlation
  - "Wine" and "Acl" (Alkalinity) have the highest correlation (.517859)
  - "Wine" and "Nonflavonoid Phenols" also exhibit correlation (.489109)

```
wine.corr(method='pearson')
```

	Wine	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue	OD	Proline
Wine	1.000000	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.499130	0.265668	-0.617369	-0.788230	-0.633717
Alcohol	-0.328222	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.136698	0.546364	-0.071747	0.072343	0.643720
Malic.acid	0.437776	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.220746	0.248985	-0.561296	-0.368710	-0.192011
Ash	-0.049643	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.009652	0.258887	-0.074667	0.003911	0.223626
Acl	0.517859	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.197327	0.018732	-0.273955	-0.276769	-0.440597
Mg	-0.209179	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.236441	0.199950	0.055398	0.066004	0.393351
Phenols	-0.719163	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.612413	-0.055136	0.433681	0.699949	0.498115
Flavanoids	-0.847498	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.652692	-0.172379	0.543479	0.787194	0.494193
Nonflavanoid.phenols	0.489109	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.365845	0.139057	-0.262640	-0.503270	-0.311385
Proanth	-0.499130	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1.000000	-0.025250	0.295544	0.519067	0.330417
Color.int	0.265668	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0.025250	1.000000	-0.521813	-0.428815	0.316100
Hue	-0.617369	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0.295544	-0.521813	1.000000	0.565468	0.236183
OD	-0.788230	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0.519067	-0.428815	0.565468	1.000000	0.312761
Proline	-0.633717	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0.330417	0.316100	0.236183	0.312761	1.000000

Alkalinity Averages

```
wine1 = wine[wine['Wine'] == 1]
```

```
average1 = wine1['Acl'].mean()
```

```
print(average1)
```

```
17.037288135593222
```

```
wine2 = wine[wine['Wine'] == 2]
```

```
average2 = wine2['Acl'].mean()
```

```
print(average2)
```

```
20.238028169014086
```

```
wine3 = wine[wine['Wine'] == 3]
```

```
average3 = wine3['Acl'].mean()
```

```
print(average3)
```

```
21.416666666666668
```

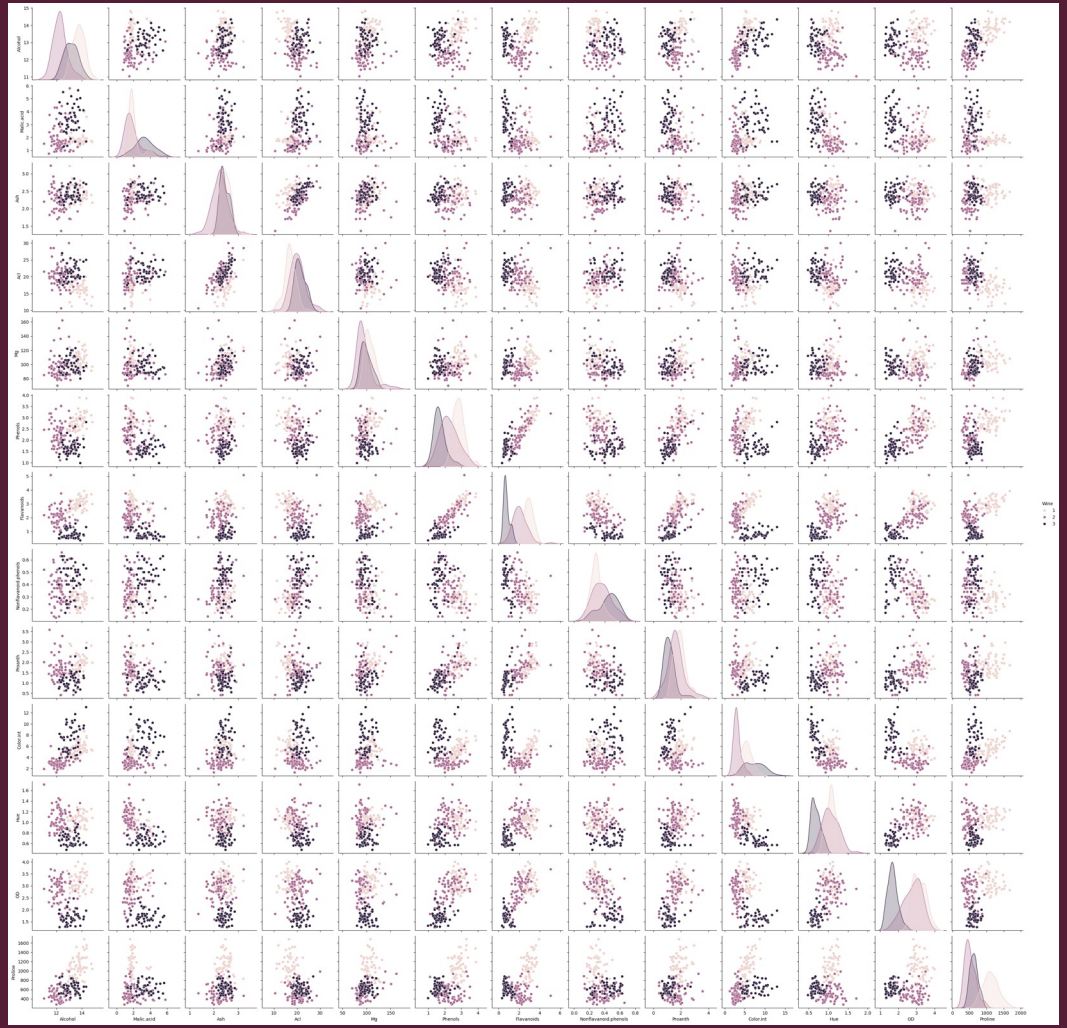
What this data means...

- Separated each wine by number (1-3) to find individual average
- Averages for alkalinity for the three origins are (slightly) different
- It can be used to determine location of each wine

Pair Plot Analysis

- Visualize relationship between each type of column variable
- Used to find association between multiple variables at the same time

Colors represent the three types of wines



Sorting Based on Conditions

- Created conditions for variable '*rating*' from existing '*Acl*' variable
  - Low: <=17, Average: 18-20, High: >=21
  - Analysis of true averages while eliminating outliers
- In most variables, the high/low/average was representative of the three wine cultivators

```
conditions = [
    (wine['Acl'] >= 21),
    (wine['Acl'] <= 17)
]
rating = ['high', 'low']
wine['rating'] = np.select(conditions, rating, default='average')
wine.rating.value_counts()
```

```
rating
average    75
high       60
low        43
Name: count, dtype: int64
```

```
wine.groupby('rating').mean()
```

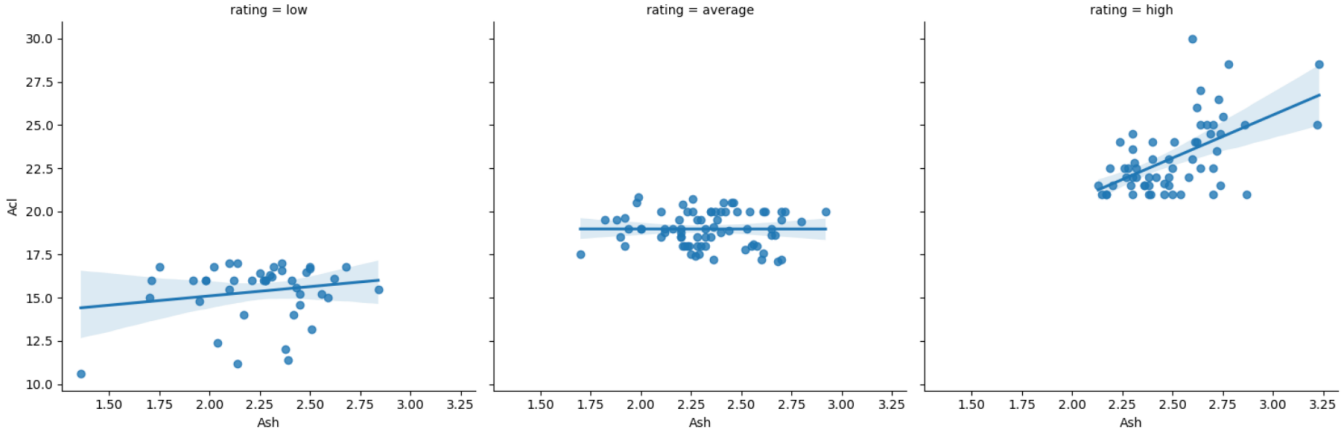
rating	Wine	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue	OD	Proline
average	1.960000	12.918667	2.316800	2.328000	18.981333	99.413333	2.213067	1.955200	0.359733	1.582533	4.969867	0.973813	2.604667	757.560000
high	2.400000	12.726000	2.778667	2.500500	23.091667	96.983333	2.089833	1.627000	0.412833	1.462667	5.074833	0.871000	2.358500	569.216667
low	1.255814	13.526744	1.753256	2.246744	15.372093	104.162791	2.724651	2.719767	0.294419	1.784419	5.188605	1.049535	2.977209	976.209302

Linear Regression

- Linear regression between ‘Ash’, ‘Acl’ (Alkalinity of Ash) & ’rating’ of red wine
- The Ash and Acl stay fairly close to the trendline because they are dependent on each other
- Average rating wines for ‘Acl’ remains almost constant irrespective of ‘Ash’
- High rating wines Ash increases as Acl increases

```
sns.lmplot(x = "Ash", y = "Acl", col = "rating", data = wine)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f194a9d3b90>
```



Key Takeaways

Alkalinity can be used to determine a wines origin because...

Ash and Alkalinity are the most correlated

The three cultivators wine's have different average alkalinities

Alkalinity is a "well behaved" variable

Questions?

GitHub: melanieann9114