conclusions_query

October 18, 2017

1 Drawing Conclusions Using Query

```
In [1]: # Load `winequality_edited.csv`
        import pandas as pd

        df = pd.read_csv('winequality_edited.csv')
        df.head()
```

```
Out[1]:    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 acidity_levels
        0      9.4        5   RED            low
        1      9.8        5   RED       med-high
        2      9.8        5   RED        med-low
        3      9.8        6   RED       med-high
        4      9.4        5   RED            low
```

1.0.1 Do wines with higher alcoholic content receive better ratings?

```
In [2]: # get the median amount of alcohol content
        alc_median = df["alcohol"].median()
```

```
In [3]: # select samples with alcohol content less than the median
        low_alcohol = df.query('alcohol < @alc_median')
        # select samples with alcohol content greater than or equal to the median
        high_alcohol = df.query('alcohol >= @alc_median')

        # ensure these queries included each sample exactly once
        num_samples = df.shape[0]
        num_samples == low_alcohol['quality'].count() + high_alcohol['quality'].count() # should
```

```
Out[3]: True
```

```
In [5]: # get mean quality rating for the low alcohol and high alcohol groups
        qlty_mean_low_alcohol = low_alcohol["quality"].mean()
        qlty_mean_high_alcohol = high_alcohol["quality"].mean()

        print("mean quality, low alcohol: {}".format(qlty_mean_low_alcohol))
        print("mean quality, high alcohol: {}".format(qlty_mean_high_alcohol))
```

```
mean quality, low alcohol: 5.475920679886686
mean quality, high alcohol: 6.146084337349397
```

1.0.2 Do sweeter wines receive better ratings?

```
In [6]: # get the median amount of residual sugar
        resid_sugar_median = df["residual_sugar"].median()
        print(resid_sugar_median)
```

```
3.0
```

```
In [7]: # select samples with residual sugar less than the median
        low_sugar = df.query('residual_sugar < @resid_sugar_median')

        # select samples with residual sugar greater than or equal to the median
        high_sugar = df.query('residual_sugar >= @resid_sugar_median')

        # ensure these queries included each sample exactly once
        num_samples == low_sugar['quality'].count() + high_sugar['quality'].count() # should be
```

```
Out[7]: True
```

```
In [8]: # get mean quality rating for the low sugar and high sugar groups
        qlty_mean_low_sugar = low_sugar["quality"].mean()
        qlty_mean_high_sugar = high_sugar["quality"].mean()

        print("mean quality, low sugar: {}".format(qlty_mean_low_sugar))
        print("mean quality, high sugar: {}".format(qlty_mean_high_sugar))
```

```
mean quality, low sugar: 5.808800743724822
mean quality, high sugar: 5.82782874617737
```

```
In [ ]:
```